# Attention Mechanism-Based Improvement of Stacked Surface Wave Cross-Correlation From High-Frequency Ambient Noise

Shufan Hu, *Member, IEEE*, Huilin Zhou, Laura Valentina Socco, and Yonghui Zhao

***Abstract*— The cross-correlation of high-frequency ambient noise (>1 Hz) is usually interpreted as the empirical Green's function between two stations and used for imaging the near surface. However, high-frequency ambient noise mainly originates from human activities with nonuniform distributions, which may lead to spurious arrival in cross-correlation and bias the analysis of surface waves. Here, we develop an algorithm for improving high-frequency surface wave cross-correlation using an attention mechanism-based neural network, CCformer. The CCformer takes two-station cross-correlations of different time segments as input. Instead of directly producing an improved cross-correlation, the CCformer integrates the process of stacking individual cross-correlations to enhance its explainability. By identifying coherent information between each segment and generating stacking weights, the CCformer improves the desired coherent signals and attenuates spurious and incoherent noises, ultimately resulting in a well-stacked cross-correlation. After training with a synthetic dataset of 200 000 labeled samples, the CCformer presents a good ability to improve the quality of stacked cross-correlation for a synthetic noise-added test dataset with dispersion, source distribution, and acquisition parameters different from the training dataset. The dispersion spectrum of the improved cross-correlation is more continuous than the results of linear stack (LS) and phase-weighted stack, and the spectral maxima agree with the theoretical dispersion curve. Moreover, a real dataset acquired from a test site also indicates the generalizability of CCformer for laterally varying media according to the symmetry of improved cross-correlation, dispersion spectrum maxima consistent with that of active data, and inversion results validated by known targets. Therefore, the proposed algorithm provides a practical solution for automatically extracting effective surface wave signals from high-frequency ambient noise.**

***Index Terms*— Ambient noise, attention mechanism, cross-correlation, surface wave.**

Received 31 December 2024; revised 15 April 2025; accepted 22 May 2025. Date of publication 29 May 2025; date of current version 10 June 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 42304158 and Grant 42074177 and in part by the Natural Science Foundation of Jiangxi Province under Grant 20232BAB212005. *(Corresponding authors: Shufan Hu; Yonghui Zhao.)*

Shufan Hu is with the School of Mathematics and Computer Sciences, Nanchang University, Nanchang 330031, China (e-mail: shufanhu@ncu.edu.cn).

Huilin Zhou is with the School of Information Engineering, Nanchang University, Nanchang 330031, China.

Laura Valentina Socco is with the Department of Geoscience and Engineering, Delft University of Technology, 2628 CN Delft, The Netherlands, and also with the Department of Environment, Land and Infrastructure Engineering (DIATI), Politecnico di Torino, 10129 Turin, Italy.

Yonghui Zhao is with the School of Ocean and Earth Science, Tongji University, Shanghai 200092, China (e-mail: zhaoyh@tongji.edu.cn).

Digital Object Identifier 10.1109/TGRS.2025.3574957

## I. Introduction

SEISMIC ambient noise, also known as microtremor or ambient vibration, is a seismic wavefield originating from human activities and natural phenomena [1]. As ambient noise is dominated by surface waves [2], its records at two stations can be cross-correlated to retrieve the empirical Green's function (EGF) with the assumption of diffuse wavefield or equipartition of the Earth's modes [3], thereby processing to infer relevant medium properties at different scales from the upper mantle to the near surface [4], [5], [6], [7]. However, the ambient noise field does not actually meet the assumption at the near-surface scale where the high-frequency (>1 Hz) ambient noise of interest mainly comes from spatially nonuniformly distributed human activities. Instead, the theory of stationary phase approximation relaxes the requirement [8], [9], and the surface wave EGF is considered to be retrieved by constructive interference of waves emitted by noise sources in stationary phase regions. Nevertheless, the presence of dominant noise sources out of stationary phase regions can corrupt the EGF with spurious arrivals [10], [11].

In practice, the ambient noise acquired for hours/days is generally subdivided into several short-duration segments, and the cross-correlation of all segments is calculated and linearly stacked to recover the EGF. Within this framework, the researchers proposed several strategies, including nonlinear stacking [12], [13], segment selection [14], [15], and denoising [16], [17], to improve the quality of stacked cross-correlations. A representative nonlinear stacking method is the phase-weighted stack (PWS) [13], which takes instantaneous phase coherence as the stacking weight to improve the quality of linearly stacked cross-correlations. Inspired by this idea, Liu et al. [18] extended the PWS to enhance the cross-correlations of the linear array by considering spatial coherency. For segment selection, it usually uses an indicator, such as the signal-to-noise ratio [14] or noise source distribution analyzed by beamforming [15], [19], to evaluate whether the cross-correlation of a time segment should be considered in stacking. Therefore, this method can be regarded as a stacking strategy where the weight of the kept segment is one, while the weight of the rejected segment is zero. Finally, denoising strategies typically use techniques, such as time–frequency filtering [20] or singular value decomposition [21], to remove noise in individual unstacked cross-correlation. It can be applied to cross-correlation of each time segment separately

or to cross-correlations of all time segments simultaneously by considering their coherence using a Wiener filter [16]. As we can see, most of the methods mentioned above leverage stacking and cross-correlation coherence between different time segments to improve the quality of final stacked cross-correlation, while spurious arrivals caused by continuous noise sources out of stationary phase regions may also present coherence, which could be the case for high-frequency ambient noise.

Machine learning, imitating the way humans learn, provides the opportunity to retrieve effective information from ambient noise. Viens and Iwata [22] used a two-step method to improve the quality of correlation calculated by deconvolution for predicting long-period ground motions from subduction earthquakes. They clustered the features in the low-dimensional space reduced by the principal component analysis and selected the best correlation function from the stacked correlation corresponding to different clusters. In the field of monitoring, Viens and Houtte [23] applied a convolutional denoising autoencoder to reduce noise in the multicomponent cross-correlation for a single station. At the exploration scale, Sun and Demanet [24] trained a modified ResNet to overcome the limitations of insufficient ambient noise recordings and nonuniform source distribution for retrieving P-wave reflections; Zhao et al. [25] designed a global multiscale fusion residual shrinkage network to denoise and encrypt the passive source virtual shot records. These machine learning-based methods enhance correlation functions based on learning from data features and have achieved success across various application fields.

To exploit the possibility of machine learning (specifically, deep learning) in improving the cross-correlation of surface waves from high-frequency ambient noise, we develop an attention mechanism-based neural network and present the procedure for training it. The main contributions of our work are summarized as follows.

1) By leveraging the attention mechanism, a global representation of individual cross-correlation within a series of cross-correlations across different time segments is achieved, allowing the network to pay attention to the expected information among all segments selectively.
2) The designed network integrates the process of stacking individual shot-duration cross-correlations, making it explainable in enhancing meaningfully coherent signals and attenuating spurious and incoherent noises.
3) The method for generating a high-quality training dataset is provided by mathematically analyzing the composition of ambient noise surface wave cross-correlation, assisting the network in reaching the desired results.

The test results of synthetic and field data show its advantages in improving surface wave cross-correlation of high-frequency ambient noise much better than traditional linear stack (LS) and PWS, both in terms of waveform and dispersion spectrum, demonstrating its effectiveness and generalizability for various complex source distributions.

The remainder of this article is organized as follows. First, we introduce the methodology of the proposed algorithm, the generation of the training dataset, and the way to training. Then, we use synthetic (with dispersion, source distribution, and acquisition parameters different from the training dataset) and field data (with multimodal dispersion and real urban complex source distribution) to test the effectiveness and demonstrate the generalizability of the proposed algorithm. Finally, we analyze and discuss the improvement of the proposed algorithm.

## II. Method

In this section, we first analyze the composition of surface wave cross-correlation. Then, we illustrate the architecture of the attention mechanism-based network for improving stacked surface wave cross-correlation. Finally, we demonstrate the generation of training samples and the training process.

### A. Surface Wave Cross-Correlation of Ambient Noise

Under the far-field assumption, the surface wave displacement response of receiver $A$ according to a single point-force source $S$ is presented as [26]

$$u_A(S,\omega) = \frac{F(\omega)e^{-\alpha(\omega)r_{SA}}}{\sqrt{r_{SA}}}e^{j[\omega t_0 - k(\omega)r_{SA}]} \tag{1}$$

where $\omega$ is the angular frequency; $F(\omega)$ and $t_0$ are the source amplitude and origin time, respectively; $r_{SA}$ is the source–receiver distance; and $k(\omega)$ and $\alpha(\omega)$ are the wavenumber and attenuation coefficient related to the medium property, respectively. As in [26], here we assume a horizontally layered medium and consider only the vertical component of the Rayleigh waves.

In the case of (1), the cross-correlation of displacement response at receivers $A$ and $B$ in the frequency domain becomes

$$\begin{aligned} C_{A,B}(S,\omega) &= u_A(S,\omega)u_B^*(S,\omega) \\ &= \frac{|F(\omega)|^2 e^{-\alpha(\omega)(r_{SA}+r_{SB})}}{\sqrt{r_{SA}r_{SB}}}e^{-jk(\omega)(r_{SA}-r_{SB})} \end{aligned} \tag{2}$$

where $r_{SB}$ is the source–receiver distance for receiver $B$ (as shown in Fig. 1). For the source located in stationary phase regions, it yields $|r_{SA} - r_{SB}| \to |r_{AB}|$, which allows us to measure the phase velocity from the phase term of cross-correlation directly, given the distance between two receivers $r_{AB}$ and the relationship $c(\omega) = \omega/k(\omega)$. Otherwise, the precise source position is necessary (but this is generally not available for ambient noise) for dispersion measurement. Ignoring the source position in the dispersion measurement (i.e., supposing that $|r_{SA} - r_{SB}|$ equals $|r_{AB}|$) could lead to underestimating the wavenumber (and, therefore, overestimating the phase velocity), as $k(\omega)|r_{SA} - r_{SB}|/|r_{AB}| < k(\omega)$ in the case of $|r_{SA} - r_{SB}| < |r_{AB}|$.

In practical preprocessing of ambient noise data, the long record of ambient noise is generally subdivided into $n$ segments. For the $i$th segment, the short-duration recording $u^i(\omega)$ can be considered as the summation of displacement corresponding to noise sources in stationary phase regions $u_s^i(\omega)$

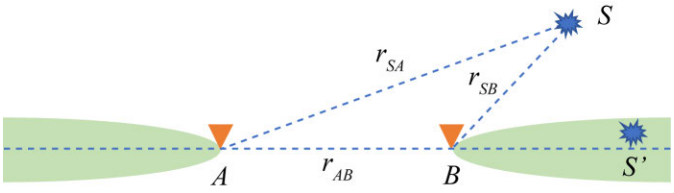


Fig. 1. Definition of geometric variables for waves propagating from a source at $S$ to two receivers. Green areas indicate stationary phase regions (modified from [8]).

and displacement related to the noise sources of nonstationary phase regions $u_n^i(\omega)$, i.e., $u^i(\omega) = u_s^i(\omega) + u_n^i(\omega)$. Then, the cross-correlation of the $i$th segment can be presented as

$$\begin{aligned} C_{A,B}^i(\omega) &= u_A^i(\omega) u_B^{i*}(\omega) \\ &= \left[u_{A,s}^i(\omega) + u_{A,n}^i(\omega)\right]\left[u_{B,s}^{i*}(\omega) + u_{B,n}^{i*}(\omega)\right] \\ &= u_{A,s}^i(\omega) u_{B,s}^{i*}(\omega) + N_i \\ &= A_i e^{-jk(\omega)r_{AB}} + N_i \end{aligned} \tag{3}$$

where $A_i$ is the amplitude term of $u_{A,s}^i(\omega) u_{B,s}^{i*}(\omega)$, which is the cross-correlation corresponding to stationary phase region sources; and $N_i$ represents the noise (i.e., spurious arrivals) in the sense of surface wave dispersion analysis, which is the summation of cross-correlation related to nonstationary phase region sources and the cross-terms

$$\begin{aligned} N_i &= u_{A,n}^i(\omega) u_{B,n}^{i*}(\omega) \\ &\quad + u_{A,s}^i(\omega) u_{B,n}^{i*}(\omega) + u_{A,n}^i(\omega) u_{B,s}^{i*}(\omega). \end{aligned} \tag{4}$$

The cross-correlations of different sources are omitted in (3) or they can be included in $N_i$.

For all $n$ segments, the linearly stacked cross-correlation is as follows:

$$\langle C_{A,B}(\omega) \rangle = \sum_{i=1}^{n} C_{A,B}^i(\omega) = \sum_{i=1}^{n} A_i e^{-jk(\omega)r_{AB}} + \sum_{i=1}^{n} N_i. \tag{5}$$

In cases of isotropic noise sources, the influence of the noise term $\sum_{i=1}^{n} N_i$ is minimized according to the destructive interference of waves originating from the nonstationary phase regions, and the surface waves are predominance in the cross-correlation waveforms by the constructive interference of waves from the stationary phase regions. However, if the sources in nonstationary phase regions are dominant, spurious arrivals can corrupt the effective surface wave signals, making it challenging to analyze surface wave dispersion accurately.

### B. Network Architecture

To improve the effective surface wave signals in stacked cross-correlation, we design the neural network (cross-correlation Transformer, CCformer), as shown in Fig. 2. The network receives two-station cross-correlation of all time segments as input. The input is linearly embedded, with no positional encoding added since the order of individual cross-correlations is unnecessary for extracting the desired coherent information from all of them. Then, the embedded vectors are fed to the Transformer encoder, which consists of $N$ identical layers. Each layer comprises multiheaded self-attention and feedforward blocks, with residual connection and layer normalization applied after both blocks. The latent vector size $d$ and the hidden size $d_{ff}$ of the feedforward block are the same for all Transformer encoder layers. In addition, we use the Gaussian error linear unit (GELU) as the activation function in the feedforward block. After the process of the Transformer encoder, the latent vector is passed to a single linear layer with a sigmoid activation function, designed to perform the linear transformation of the latent vector to match the number of samples in cross-correlation and produce stacking weights between 0 and 1. Finally, the original time-segmented cross-correlation input is multiplied by the corresponding stacking weight, and a time-domain summation (i.e., stacking) is implemented to output the single improved stacked cross-correlation.

Given the input $\mathbf{X} = [\boldsymbol{x}_1; \boldsymbol{x}_2; \ldots; \boldsymbol{x}_n]$, where $\boldsymbol{x}_i \in \mathbb{R}^{1\times ns}$ is a row vector containing the two-station time-domain cross-correlation with $ns$ samples for the $i$th segment, the $j$th sample of the improved stacked cross-correlation $\hat{y}_j$ output from the network is calculated as follows:

$$\mathbf{Z}_0 = \mathbf{X}\mathbf{E} \tag{6}$$

$$\mathbf{Z}_l' = \mathrm{LN}(\mathrm{MSA}(\mathbf{Z}_{l-1}) + \mathbf{Z}_{l-1}), \quad l = 1, 2, \ldots, N \tag{7}$$

$$\mathbf{Z}_l = \mathrm{LN}\left(\mathrm{GELU}\left(\mathbf{Z}_l'\mathbf{W}_l^{(1)} + \boldsymbol{b}_l^{(1)}\right)\mathbf{W}_l^{(2)} + \boldsymbol{b}_l^{(2)} + \mathbf{Z}_l'\right) \quad l = 1, 2, \ldots, N \tag{8}$$

$$\hat{y}_j = \sum_i (\mathbf{X} \cdot \mathrm{sigmoid}(\mathbf{Z}_N\mathbf{W} + \mathbf{b}))_{i,j} \tag{9}$$

where $\mathbf{E} \in \mathbb{R}^{ns\times d}$ is the parameter matrix in linear embedding; LN and MSA indicate layer normalization and multiheaded self-attention [27], respectively; $l = 1, 2, \ldots, N$ represents the $l$th layer in the Transformer encoder; $\mathbf{W}_l^{(1)} \in \mathbb{R}^{d\times d_{ff}}$ and $\mathbf{W}_l^{(2)} \in \mathbb{R}^{d_{ff}\times d}$ are the weights in the feedforward block; $\mathbf{b}_l^{(1)} \in \mathbb{R}^{1\times d_{ff}}$ and $\mathbf{b}_l^{(2)} \in \mathbb{R}^{1\times d}$ are the corresponding biases; and $\mathbf{W} \in \mathbb{R}^{d\times ns}$ and $\mathbf{b} \in \mathbb{R}^{1\times ns}$ are the weight and bias for the final linear layer, respectively.

For the MSA block, it parallelly performs $h$ attention functions. Each attention function (here, it is the scaled dot-product attention) produces outputs by

$$\mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right) \tag{10}$$

where $\mathbf{Q} \in \mathbb{R}^{n\times d_k}$, $\mathbf{K} \in \mathbb{R}^{n\times d_k}$, and $\mathbf{V} \in \mathbb{R}^{n\times d_v}$ are the matrices linearly projected from the input of the MSA block. As in [27], we use $d_k = d_v = d/h$.

According to (10), the attention mechanism focuses on the parts of the input most relevant to the desired output by measuring similarity, making it inherently well suited for extracting coherent information from cross-correlations of all segments. Moreover, it achieves a global representation of individual cross-correlation within a series of cross-correlations across different time segments. This benefit enables the network to understand the importance of specific individual correlations without being affected by their order in the input sequence. With the integration of the stacking process, the network is considered easier to train to produce a high stacking weight for coherently effective surface wave signals and a low stacking weight for spurious and incoherent noises and, therefore, an improved stacked surface wave cross-correlation.

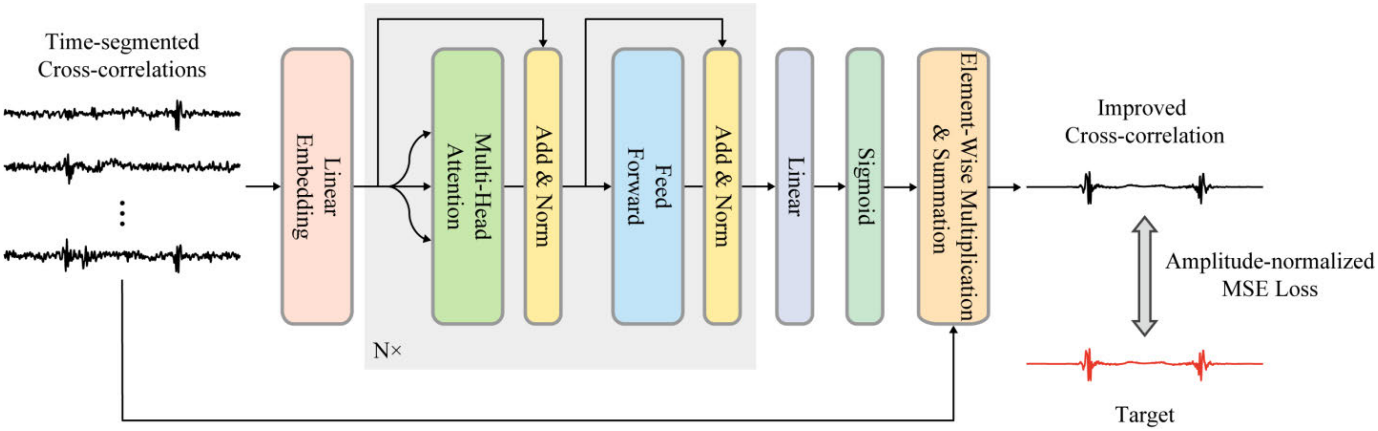


Fig. 2. Architecture of the network for improving surface wave cross-correlation.

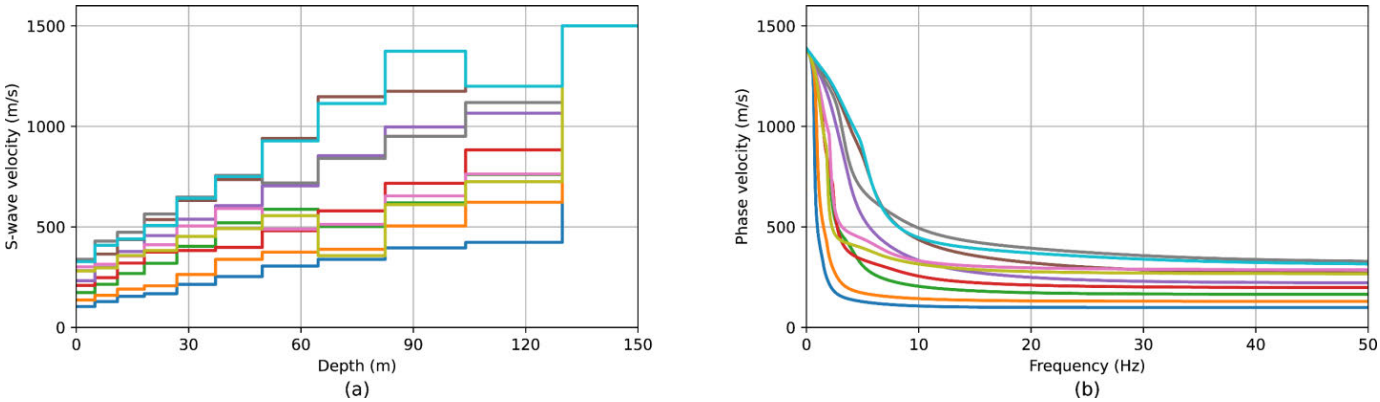


Fig. 3. (a) Ten randomly selected samples of the generated elastic models and (b) corresponding dispersion curves.

### C. Dataset

With the input being the two-station cross-correlation of all segments and the output being the single improved stacked two-station cross-correlation, we use numerical simulation to generate the dataset for training the network. We first establish 200 000 1-D elastic models using the constrained Markov decision method [28]. Each elastic model consists of ten layers with a half-space. The top layer has an S-wave velocity varying from 100 to 350 m/s and a thickness of 5 m; the S-wave velocity of half-space is fixed to 1500 m/s; for other layers, its thickness is 1.2 times that of the upper layer, and the S-wave velocity is determined according to the upper layer with different possibilities to increase or decrease. After generating the elastic models, we calculate the corresponding phase–velocity dispersion curve of the fundamental mode using the transfer matrix method [29], [30]. Fig. 3 shows ten randomly selected elastic models and their dispersion curves. Note that using the specific stratum and fundamental dispersion in dataset generation does not limit the network because the task is to reduce the influence of nonuniformly distributed noise sources, and the main point is to generate a dataset with abundant diversity.

We then use the method in [26] to simulate ambient noise records at two stations for each elastic model. In the simulation, the noise source signature is a random sequence constructed according to the method in [31], with a duration of fewer than 5 s, a maximum amplitude in the time domain satisfying the normal distribution $N(0, 1)$, and a maximum frequency between 1 and 50 Hz. We place 1000 noise sources randomly distributed at the polar coordinate with a radius between 1 and 2 km. Under this setting, the spatial distribution of noise sources is nearly uniform along the azimuth; however, the source distribution is indeed nonuniform in terms of the amplitude of each frequency component since the maximum frequency of each noise source is different. For the distance between the two stations, we set it to vary from 1 to 200 m. After simulation with a recording time of 30 min and a sampling interval of 4 ms, we obtain the two-station ambient noise recordings for each elastic model. Fig. 4 presents the spatial distribution of noise sources, the source signature of 20 noise sources, and the simulated ambient noise, for one of the generated elastic models.

For each simulated two-station ambient noise recording, we preprocess it to generate the input of the network, following the procedure described by Bensen et al. [32].

1) Remove the mean and trend of the signals recorded at each station.
2) Apply a bandpass filter with cutoff frequencies of 0.05 and 30 Hz (we limit the cutoff frequency to 30 Hz to avoid tending to retrieve relevant information of higher frequencies but rare sources with these frequencies located in stationary phase regions).

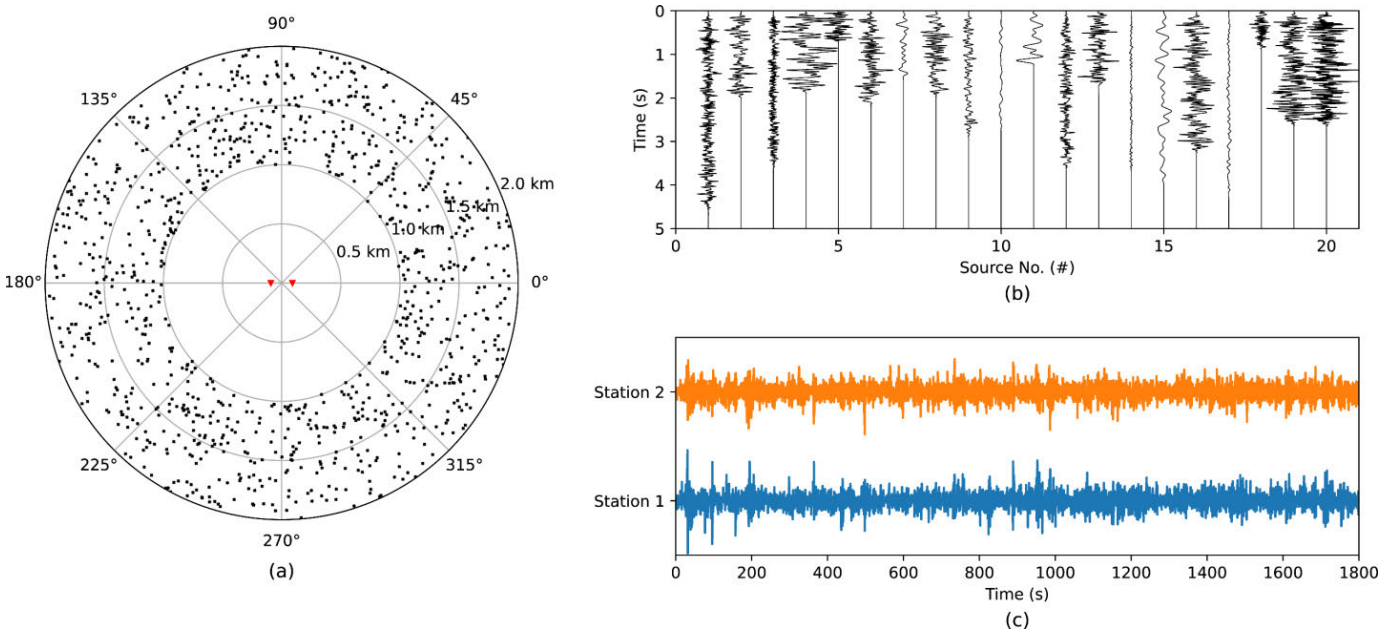


Fig. 4. Ambient noise simulation for one of the generated elastic models. (a) Spatial distribution of noise sources, (b) source signature of 20 noise sources, and (c) simulated ambient noise.

3) Cut the continuous recordings into segments with a length of 30 s and an overlap of 15 s.
4) Employ the running absolute mean normalization and spectral whitening to each segment.
5) Compute the cross-coherence $H^i_{A,B}(t)$ (the cross-correlation discarding amplitude information) [33] between two stations for each segment as

$$H^i_{A,B}(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{u^i_A(\omega)u^{i*}_B(\omega)}{\left|u^i_A(\omega)\right|\left|u^i_B(\omega)\right|} e^{j\omega t} d\omega. \quad (11)$$

For the ground truth $y$, it is calculated by the reverse Fourier transform as

$$y(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-jk(\omega)r_{AB}} e^{j\omega t} d\omega \quad (12)$$

according to the phase information of the cross-correlation corresponding to the stationary phase region sources, as in (3). Note that (12) does not introduce a $\pi/4$ phase shift, as we only expect to extract the relevant information corresponding to sources located in stationary phase regions rather than transform the cross-correlations into a homogenous source condition [34]. Moreover, only the frequency component within the cutoff frequencies is calculated for generating the ground truth of (12) since we apply a bandpass filter in the preprocessing to obtain the input. Fig. 5 shows an example of the input and ground truth in the dataset.

### D. Training

We set the parameter of the CCFormer similar to the base model in [35], i.e., the number of layers $N$ in the Transformer encoder is 6, the latent vector size $d$ is 256, the number of heads $h$ in the MSA block is 4, and the hidden size of the feedforward block $d_{ff}$ is 1024, resulting in a total of 5 508 573 trainable parameters. Since the magnitude of the network output (a stacked cross-correlation) can differ from that of ground truth, we define an amplitude-normalized mean squared error (mse) loss function as

$$\text{Loss} = \frac{1}{ns}\left\| \boldsymbol{y}_{\text{norm}} - \hat{\boldsymbol{y}}_{\text{norm}} \right\|_2^2 \quad (13)$$

where $\boldsymbol{y}_{\text{norm}}$ and $\hat{\boldsymbol{y}}_{\text{norm}}$ are the normalized ground truth and network output, respectively, and their element is between $-1$ and 1.

We then train the network using Pytorch on an NVIDIA GeForce RTX 4090D graphics processor. To check the convergence and avoid overfitting, we randomly select 70% of the dataset for the training and 30% as the validation dataset. The batch size is set to 128, and an AdamW optimizer [36] with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-9}$ is used. In addition, we vary the learning rate over the course of training using the warmup schedule [35]

$$lr = \alpha \cdot d^{-0.5} \cdot \min\left(k^{-0.5}, k \cdot \text{warmup}_k^{-1.5}\right) \quad (14)$$

where $k$ is the step number, $\alpha$ is a tunable scalar, and $\text{warmup}_k$ is the number of steps increasing the learning rate in the early training. Here, we use $\alpha = 0.25$ and $\text{warmup}_k = 4000$. Fig. 6 shows the training loss and validation loss, which indicate a good convergence and no overfitting is observed.

## III. Result

In this section, we use both synthetic noise-added data (with dispersion, source distribution, and acquisition parameters different from the training dataset) and field data from a test site (with multimodal dispersion and real urban complex source distribution) to evaluate the effectiveness and test the generalization of the CCformer. We also compare the improved cross-correlations and their corresponding dispersion spectrum to those of LS and PWS (a default power $v = 1$ is used throughout this work).

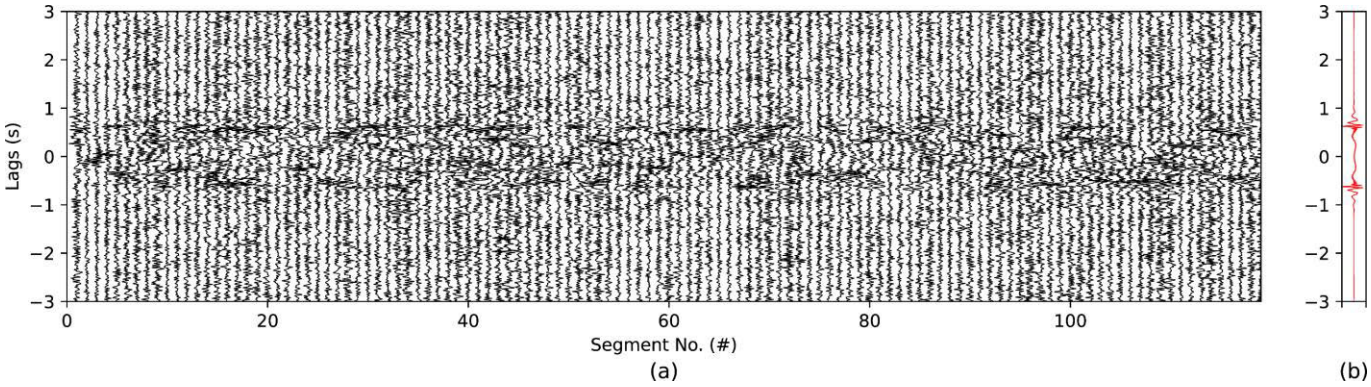


Fig. 5. Example of (a) input and (b) ground truth in the dataset.

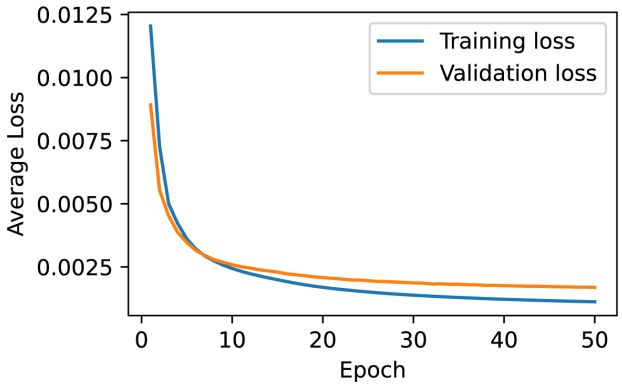


Fig. 6. Training and validation losses of the CCformer.

### A. Synthetic Data

The synthetic data are generated using an S-wave velocity model [Fig. 7(a)] similar to the inversion result of field data in [37], which is different from the models used in generating the training dataset. To simulate the ambient noise recordings, we distribute 1000 noise sources in the polar coordinate, with a radius ranging from 1 to 2 km. Of these noise sources, half of them are uniformly distributed, and the other half of them are symmetrically arranged along a specific azimuth (either $\pi/2$ or $3\pi/4$) with ranges of $\pm\pi/2$, $\pm\pi/3$, and $\pm\pi/6$ (we refer to it as nonuniformly distributed noise sources). This setup results in six different configurations for the source distribution, as illustrated by Tests 1–Test 6 in Fig. 7(b)–(g). The source duration, amplitude, and frequency used in the simulation are the same as in the training dataset generation, but a sampling interval of 5 ms (compared to 4 ms of the training dataset) is adopted to test the performance of CCformer in processing input with different acquisition parameters. We simulate 30-min ambient noise for a linear array of 41 stations with an equal spacing of 5 m. In addition, we add Gaussian noise with a mean of 0 and a standard deviation of 5% of the maximum amplitude to each simulated ambient noise data. As a result, it produces synthetic data with differences in dispersion, source distribution, and acquisition parameters compared to the training dataset, making it suitable for testing the ability of the proposed algorithm.

After simulation, we preprocess the synthetic data following the same procedure as preparing the training dataset, with the leftmost station as the virtual source. This results in unstacked cross-correlations for each time segment and each possible station pair. Then, we use the proposed CCformer to generate improved stacked cross-correlations. We also produce LS and PWS results for comparison.

Fig. 8 shows the stacked cross-correlations and their corresponding dispersion spectra for configurations with half of the sources being symmetrically distributed along the azimuth of $3\pi/4$ (i.e., Tests 1–3), using the LS, PWS, and CCformer. It indicates that, for these three cases, the cross-correlations improved by the CCformer present a better symmetry between causal and acausal branches than the results of LS and PWS. Moreover, the waveform of effective surface waves is enhanced in the CCformer results, clearly showing the dispersion characteristic as the station distance increases despite the cross-correlation of each station pair being processed separately. For the dispersion spectrum, the energy maxima in each spectrum of CCformer display good continuity along the frequency axis, and they agree with the theoretical dispersion curve at wavelengths shorter than the maximum array aperture (i.e., the length of the linear array). In contrast, although the influence of nonuniformly distributed sources on the spectrum of LS is insignificant for Tests 1 and 2, the spectrum of Test 3 produced by the LS overestimates the phase velocity at frequencies below 10 Hz. It also exhibits strong artifacts at frequencies above 25 Hz, as most of the nonuniformly distributed sources in Test 3 are out of the stationary phase regions and consequently cause considerable spurious arrivals. The PWS additionally leads to an increase in spurious arrivals for Test 3 that the dispersion spectrum shows more artifact energies at a wide frequency band and a gap at frequencies near 5 Hz. This is because spurious arrivals originating from sources with similar azimuth can also present coherence, which is noticeable by early arrivals in the causal branch of the corresponding cross-correlations for Test 3.

Fig. 9 shows the stacked cross-correlations and their corresponding dispersion spectra for configurations with half of the sources being symmetrically distributed along the azimuth of $\pi/2$ (i.e., Tests 4–6), using the LS, PWS, and CCformer.

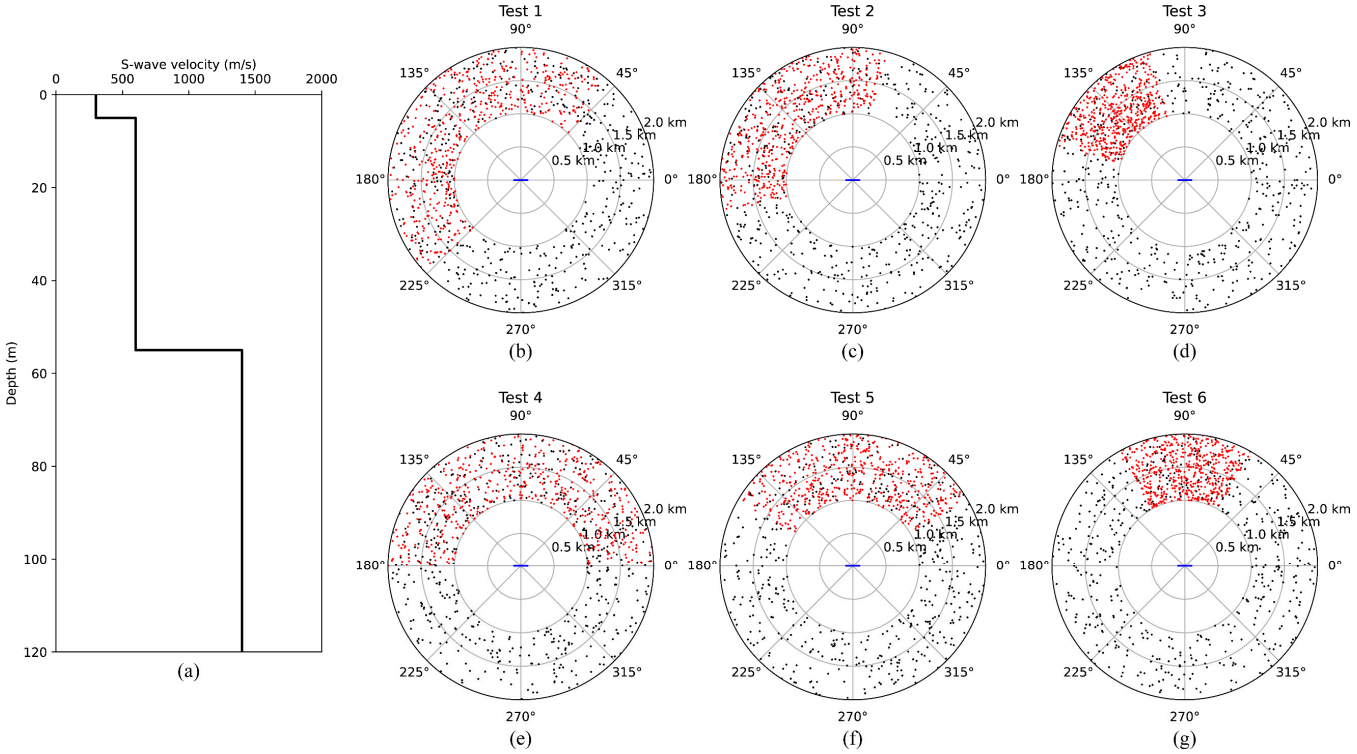


Fig. 7. (a) S-wave velocity model and (b)–(g) placements of noise sources for simulating the synthetic data. The black and red dots in (b)–(g) indicate uniformly and nonuniformly distributed noise sources, respectively. The blue line in (b)–(g) represents the linearly receiving array consisting of 41 stations.

For Test 4, where all sources can be considered uniformly distributed (the sources distributed within the azimuth of $[0, \pi]$ are essentially identical to the one of $[0, 2\pi]$ for the linear array), the spectral maxima in the dispersion spectrum for LS, PWS, and CCformer are similar and agree with the theoretical dispersion curve at wavelengths shorter than the maximum array aperture. Notably, the CCformer produces the clearest waveform of effective surface waves by attenuating possible background noise caused by cross-correlating responses from separated sources. For Tests 5 and 6, where most noise sources are out of stationary phase regions, significant spurious arrivals are present in the cross-correlation of LS and PWS. These spurious arrivals result in a dispersion spectrum containing numerous artifact energies, making it challenging to estimate phase velocity (especially for Test 6). Nevertheless, the CCformer still produces cross-correlations with high-quality waveforms of surface waves, and the corresponding dispersion spectrum clearly reflects the expected dispersion pattern consistent with the theoretical dispersion curve. Although there are some slight oscillations in the dispersion spectrum of CCformer, we consider it acceptable because, in these two cases, the nonuniformity of the noise source is very strong, and additional random noise is added to the simulated ambient noise.

According to these tests, the CCformer does not reduce the quality of stacked cross-correlations when the impact of nonuniformly distributed noise sources is slight; for cases where the influence of nonuniformly distributed noise sources is significant, it can effectively eliminate spurious arrivals in stacked cross-correlations and enhance the dispersion spectrum, producing results superior to that of LS and PWS methods.

### B. Field Data

The field data are acquired from a test site in Shanghai, China. This test site was first designed with known buried targets, including precast concrete or stone near the site, to evaluate the capabilities of the EM wave method in underground space exploration. The design depths of these targets are 3–20 m, placed by drilling a hole. However, their depths can exceed the design values due to settling within the drilling mud.

About three months after the construction of the site, we measured ambient noise by 17 × 17 (i.e., 289) Earth Pulse Ant-1C nodal seismographs (5 Hz), using the deployment of a small regular seismic network with an even spacing of 1 m along the $x$- and $y$-directions, as shown in Fig. 10. Moreover, traffic on nearby roads [Fig. 10(a)] around the site provides potential high-frequency noise sources. Although we deploy a seismic network, the purpose of the measurement is to exploit the possibility of using urban high-frequency ambient noise recorded by a linear array to estimate accurate surface wave dispersion as active data, given that the linear array is convenient to deploy in urban areas and, further, to retrieve the shallow laterally varying S-wave velocity model (mainly for Line 6 that the targets are stone and with shallowest design depth of 3 m). Therefore, we treat the seismic network as several survey lines. For each station, we recorded 24 h continuous ambient noise with a sampling rate of 2 ms

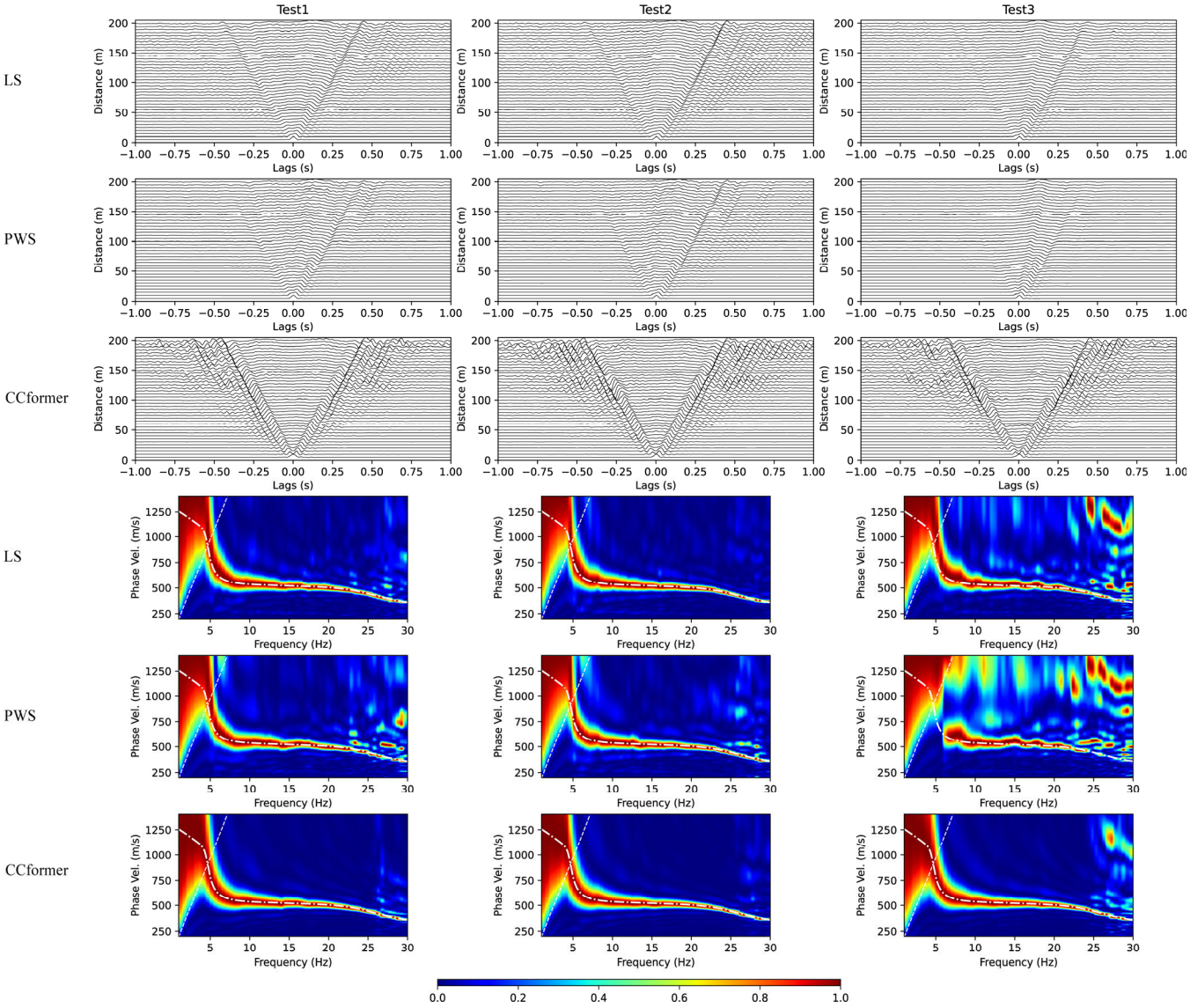


Fig. 8. Improvement results for Tests 1–3 of synthetic data. The first to third rows depict the cross-correlations produced by LS, PWS, and CCformer, respectively. The fourth to sixth rows represent the corresponding dispersion spectrum for LS, PWS, and CCformer results, respectively. The white dotted-dashed line in the dispersion spectrum indicates the theoretical dispersion curve, and the white dashed line in the dispersion spectrum demonstrates the wavelength equal to the maximum array aperture. Note that all wiggle plots of the cross-correlations are on the same scale.

(compared to 4 ms of the training dataset). At the same time, we acquired active data by shooting near the first station of each line [an offset of about 0.5 m, Fig. 10(c)]. This allows us to check the quality of cross-correlations retrieved from ambient noise by comparing the dispersion spectra with the ones of active data.

First, we use the ambient noise data of Lines 1, 9, and 17 (i.e., the first, middle, and last line) to evaluate the capability of estimating phase velocity for the urban high-frequency ambient noise recorded by the linear array. We preprocess the data using the same procedure in generating the training dataset, taking the first station as the virtual source. After obtaining unstacked cross-correlation for each time segment and each possible station pair, we produce the stacked cross-correlation using LS, PWS, and CCformer, as shown in Fig. 11. It indicates that, for each line, the CCformer produces the cross-correlations with the best symmetry compared to the results of LS and PWS, demonstrating the advantage of the CCformer in reducing the influence of the nonuniformly distributed noise sources. Fig. 12 depicts the dispersion spectra corresponding to cross-correlations in Fig. 11, with comparisons to the spectral maxima on the dispersion spectrum of active data. It should be noted that there are lateral variations along Line 9 according to the known buried targets [Fig. 10(c)], and the media under Lines 1 and 17 are assumed to be laterally homogenous since no targets are designed along the two lines. For Line 1, the energy on the dispersion spectrum of CCformer coincides with the spectral maxima of the active data at wavelengths shorter than the maximum array aperture. In contrast, there is a gap near frequencies of 25 Hz in the high mode of the LS spectrum, and the PWS spectrum exhibits crosstalk between different modes at the frequency of about 22 Hz. For Line 9, despite dispersion spectra of LS and PWS exhibiting a gap of energy at frequencies near 12 Hz, the one of CCformer presents dispersive energy with good continuity along the frequency axis, and

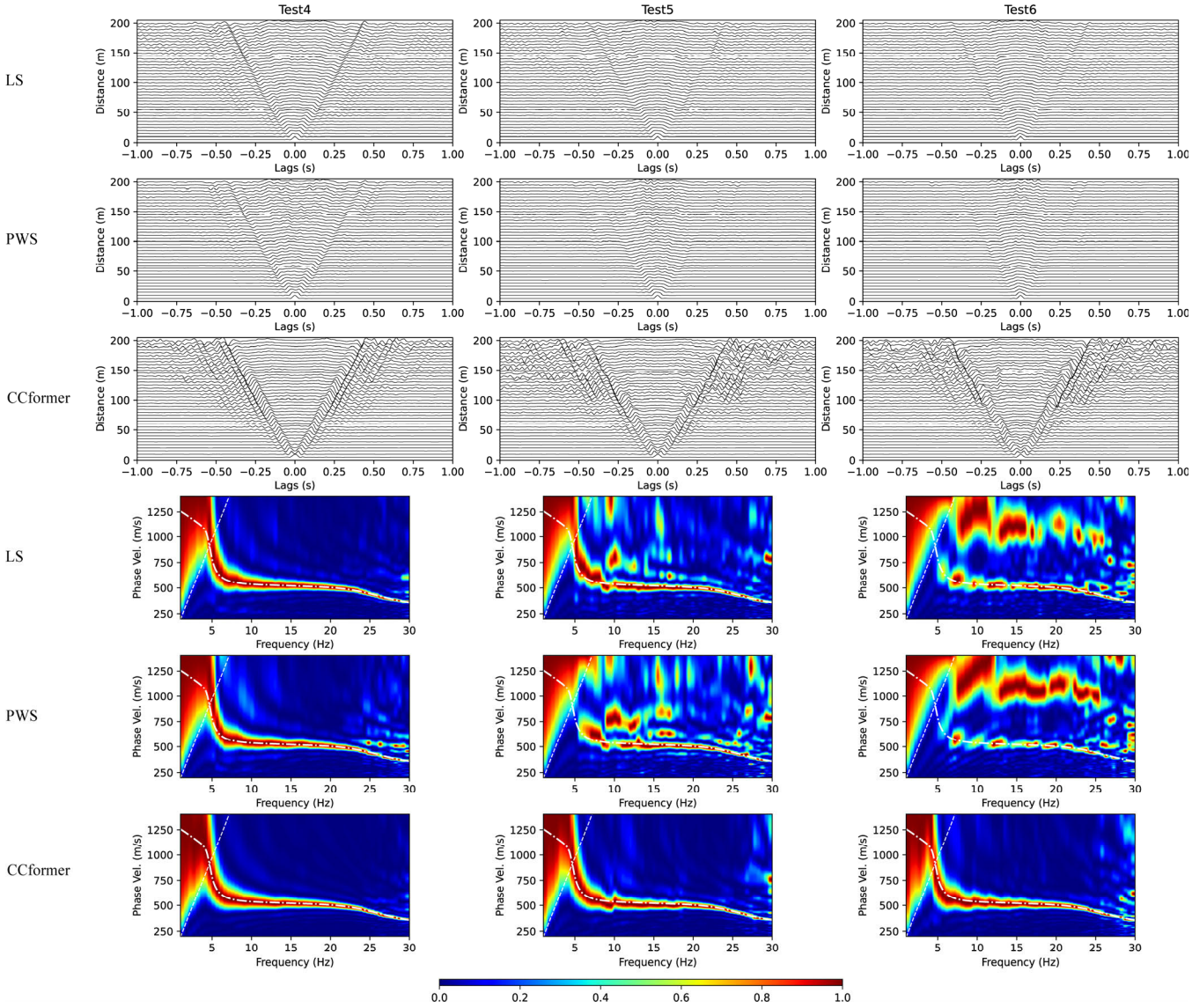


Fig. 9. Improvement results for Tests 4–6 of synthetic data. The first to third rows depict the cross-correlations produced by LS, PWS, and CCformer, respectively. The fourth to sixth rows represent the corresponding dispersion spectrum for LS, PWS, and CCformer results, respectively. The white dotted-dashed line in the dispersion spectrum indicates the theoretical dispersion curve, and the white dashed line in the dispersion spectrum demonstrates the wavelength equal to the maximum array aperture. Note that all wiggle plots of the cross-correlations are on the same scale.

the spectral maxima agree with those of active data at most frequencies. The slight deviation at frequencies lower than 12 Hz may arise from the differences in locations between active and virtual sources, as lateral variations are presented along this line [Fig. 10(c)]. Finally, for Line 17, although the dispersion spectrum of LS displays a gap of energy at frequencies near 12 Hz, both PWS and CCformer produce the dispersion spectrum with energy that agrees with the spectral maxima of active data at wavelengths shorter than the maximum array aperture. Notably, the dispersion spectrum of Line 17 from the CCformer additionally presents the advantage that more low-frequency energy related to the higher mode is retrieved, even though we consider only the fundamental mode in the training stage. These tests indicate that the CCformer effectively mitigates the adverse influence of noise sources out of stationary phase regions and improves the stacked cross-correlations, making it possible to estimate accurate surface wave dispersion from urban high-frequency ambient noise recordings of the linear array.

As the CCformer effectively improves the stacked cross-correlations of high-frequency ambient noise recorded by the linear array, we use the MWASW method [38] to retrieve the 2-D S-wave velocity model of Line 6 by inverting the dispersion data analyzed from the improved stacked cross-correlations. By taking all stations as the virtual source, we first calculate the cross-correlations of all possible station pairs in Line 6 and use the CCformer to improve them, resulting in 17 common virtual-source gathers. The use of multiple virtual sources at different locations enhances dispersion stacking, improving the signal-to-noise ratio of the dispersion spectrum and benefiting subsequent dispersion analysis. Then, we use spatial windows with multiple sizes, successively varying from 7 to 15 stations, to extract several dispersion curves along the survey line. Fig. 13(a) shows extracted

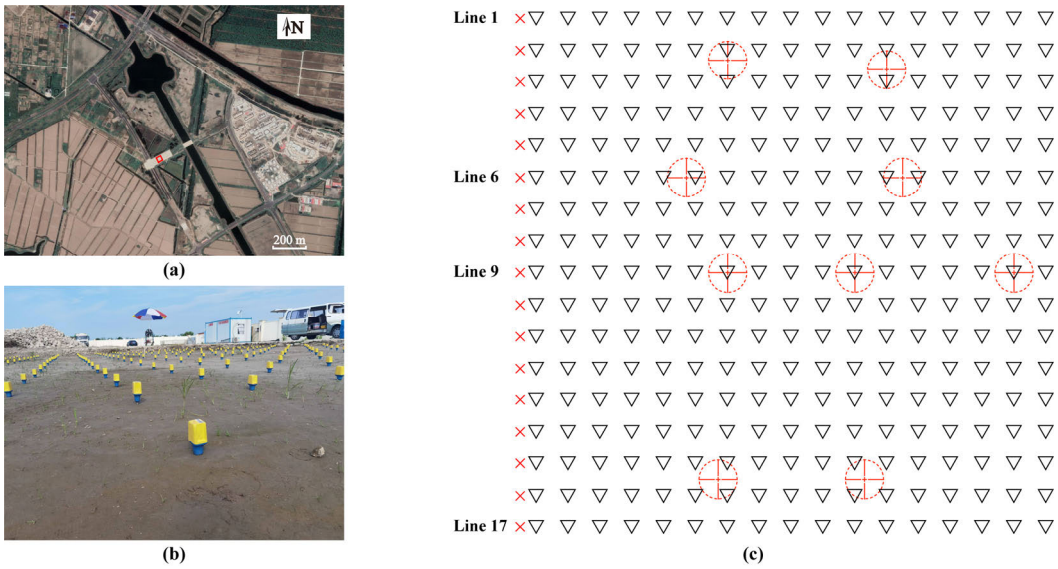


Fig. 10. (a) Satellite image and (b) photograph of the test site, along with the deployment of (c) seismic network for acquiring field data. In (c), the red dashed circles with crosses indicate the designed positions for targets, and the red crosses denote the positions of active sources.

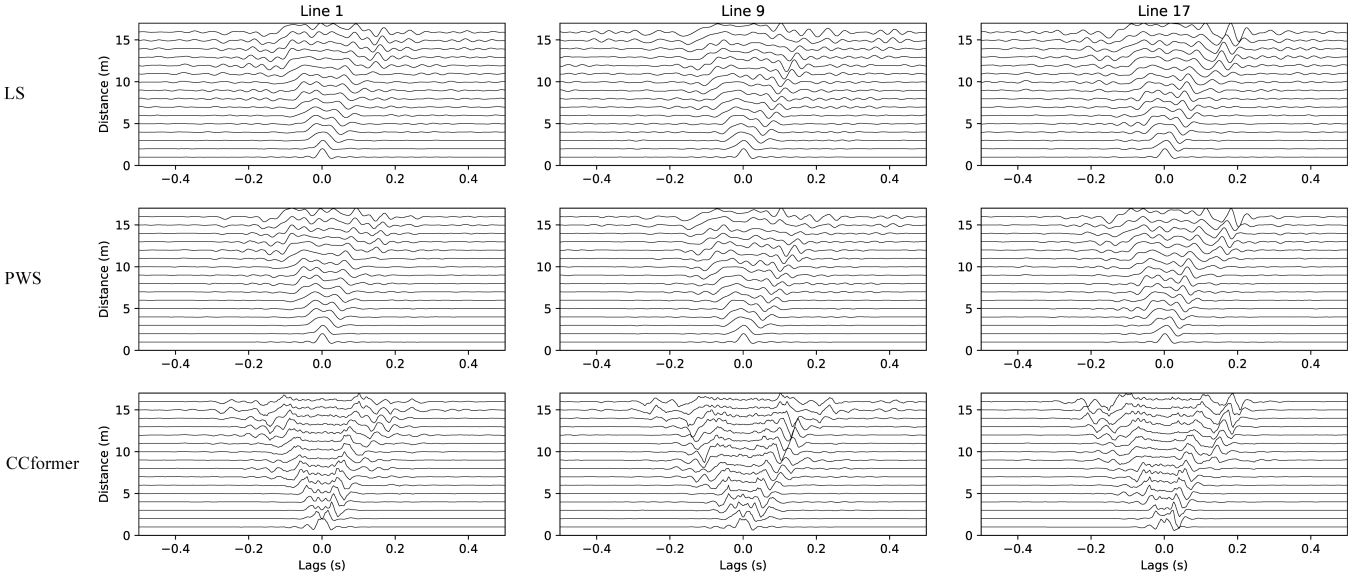


Fig. 11. Improvement results on cross-correlations for Lines 1, 9, and 17 of filed data. The first to third rows depict the cross-correlations produced by LS, PWS, and CCformer, respectively. All wiggle plots of the cross-correlations are on the same scale.

dispersion data for all spatial windows. Note that the active data for Line 6, obtained from a single shooting position, cannot yield such dispersion curves since the signal-to-noise ratio of the dispersion spectrum for the small window in far offset is too low to identify the accurate dispersion (not shown here). Finally, using the Gauss–Newton optimization and a 2-D forward algorithm, we simultaneously invert all estimated dispersion curves to obtain the 2-D S-wave velocity model, as depicted in Fig. 13(c). The retrieved S-wave velocity model highlights two high-velocity areas. The left high-velocity area corresponds well to the intended design, while the right high-velocity area is positioned slightly further to the right than planned. This difference may be attributed to errors in station deployment and site construction. Despite this, we consider the quality of the retrieved model to be acceptable since only 17 stations are used. The retrieved targets also indicate that the stacked cross-correlation produced by the CCformer has been well improved for later surface wave analysis and inversion.

## IV. Discussion

We presented an attention mechanism-based network, CCformer, to improve the stacked high-frequency ambient noise surface wave cross-correlation for estimating surface wave dispersion. The CCformer incorporates the routine procedure in traditional ambient noise processing, i.e., stacking the original individual cross-correlations of each time segment, and the novelty includes two aspects: 1) the network leverages the attention mechanism to pay attention to desired coherent information between cross-correlations of different

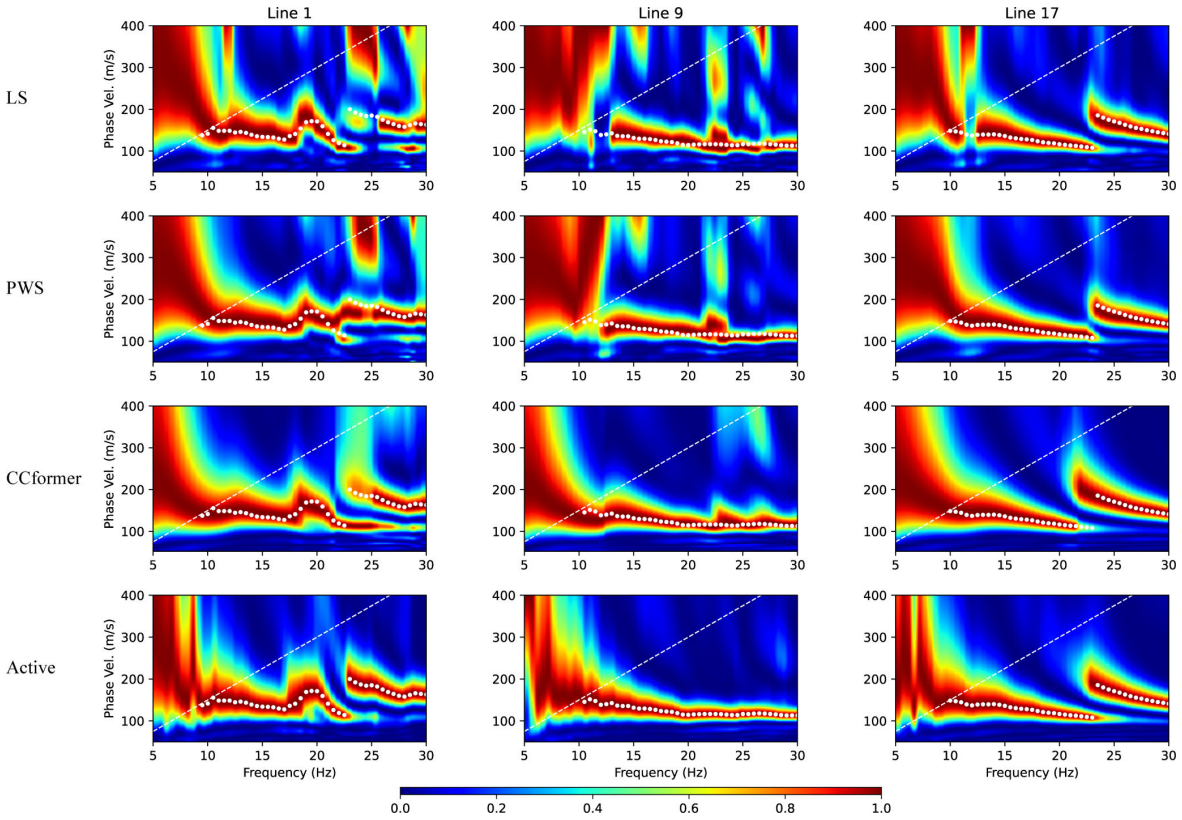


Fig. 12. Results on dispersion spectra for Lines 1, 9, and 17 of filed data. The first to fourth rows depict the dispersion spectrum for the result of LS, PWS, CCformer, and active data, respectively. The white dotted line in the dispersion spectra indicates the spectral maxima on the dispersion spectrum of active data, and the white dashed line in the dispersion spectrum demonstrates the wavelength equal to the maximum array aperture.

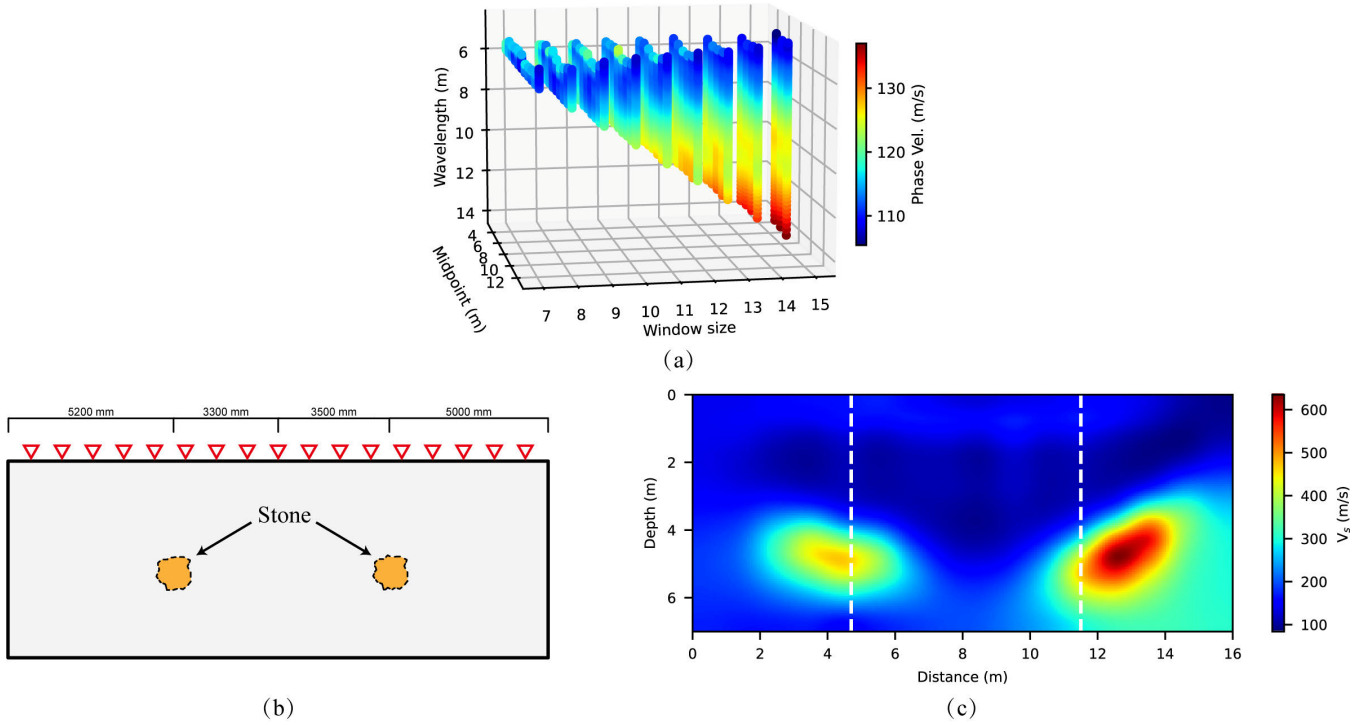


Fig. 13. (a) Extracted dispersion curves, (b) sketch of embedded targets, and (c) retrieved S-wave velocity model for Line 6 of the site.

time segments and 2) it is designed to integrate the stacking process and is trained to learn the stacking weights for enhancing effectively coherent signals and attenuating spurious and incoherent noises. As a result, the network is more straightforward to train under our carefully established training dataset, producing improved cross-correlation. Tests on

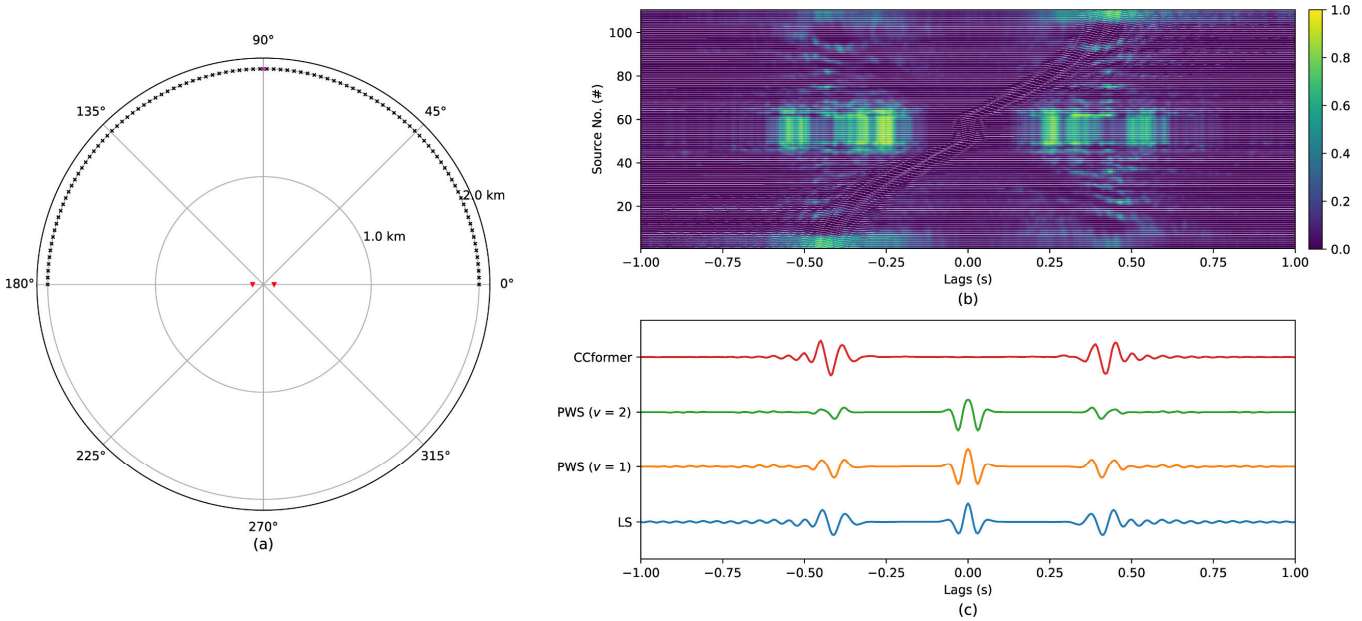


Fig. 14. Test result for checking the stacking weight produced by CCformer. (a) Placements of sources and receivers, (b) produced stacking weight overlapped with input cross-correlations, and (c) stacked cross-correlations produced by LS, PWS, and CCformer.

synthetic and field data all show the effectiveness and generalizability of the proposed CCformer.

Most traditional methods, such as nonlinear stacking and segment selection [12], [13], [14], [15], work with the idea of stacking more useful cross-correlations. This concept inspired the CCformer, making it explainable rather than operating as a opaque system. To check whether the CCformer works as it should, we carry out a test to show the stacking weight it generates during processing. In this test, we distribute 110 sources at a radius of 2 km to simulate a situation where dominant coherent out-line sources are presented in a homogeneous wavefield. Of these sources, 100 are arranged with azimuth linearly increasing from 0 to $\pi$, and the remaining ten sources are located in the azimuth of $\pi/2$, as shown in Fig. 14(a). For each source, we directly calculate the cross-correlation of two receivers (with a receiver distance of 200 m) according to the formula of (2), using a Ricker wavelet with a center frequency of 15 Hz and the same model in Fig. 7(a). The time lags of the cross-correlation are between $-1$ and 1 s. We then consider the cross-correlation of each source as the ambient noise cross-correlation of a time segment and pass them to the CCformer. During its processing, we save the intermediate product, i.e., the stacking weight between 0 and 1 output by the final linear layer with a sigmoid activation function. Fig. 14(b) depicts the stacking weight output by the final linear layer overlapped with input cross-correlations. Although the ten cross-correlations corresponding to the sources at an azimuth of $\pi/2$ are the same, the CCformer produces high stacking weight for events only in the cross-correlation related to the stationary phase region sources, i.e., the CCformer pays attention only to the coherent waveform with late arrivals despite that there is earlier event showing higher consistency. Moreover, the stacking weights exhibit a certain symmetry, indicating that the CCformer seeks to extract useful information from either the causal or acausal branch when coherent waveforms are present in one of the branches. Notably, high stacking weights (with some values almost close to 1) are also observed for the cross-correlations of sources at the azimuth near $\pi/2$, with corresponding time lags where the stacking weights are high for the cross-correlation related to stationary phase region sources. This seems to be that the CCformer tends to preserve the zeros in the cross-correlations with significant deviation from the effective event or it attempts to extract information from cross-correlation with significant arrivals near zero time lag (this could be the case that there are dominant sources out of stationary phase regions, but effective information is present in the cross-correlation). We also present the final stacked cross-correlations produced by LS, PWS, and CCformer, as shown in Fig. 14(c). It indicates that the LS does not effectively attenuate the arrival at zero time lag caused by sources at an azimuth near $\pi/2$, and the PWS worsens this issue, as this arrival remains consistent across multiple cross-correlations. In contrast, the proposed CCformer successfully attenuates it by generating a stacking weight close to zero, consequently improving the effective event. This test suggests that the working process of the CCformer is more explainable than an end-to-end opaque system, and we can use the intermediately generated stacking weight to verify the stacked cross-correlation output by the neural network.

Although long-term recordings of ambient noise, such as over months or years, are generally considered necessary to retrieve high-quality cross-correlations [11], short-term deployments (such as several minutes) are often favored for near-surface surveys due to site conditions and cost considerations [17], [18], [19], [39]. However, the field data results (Figs. 11 and 12) indicate that the recordings of 24 h may still not be sufficient to retrieve reliable cross-correlations using the traditional LS. We then use field data of Line 17 to test the performance of the CCformer in improving short-term

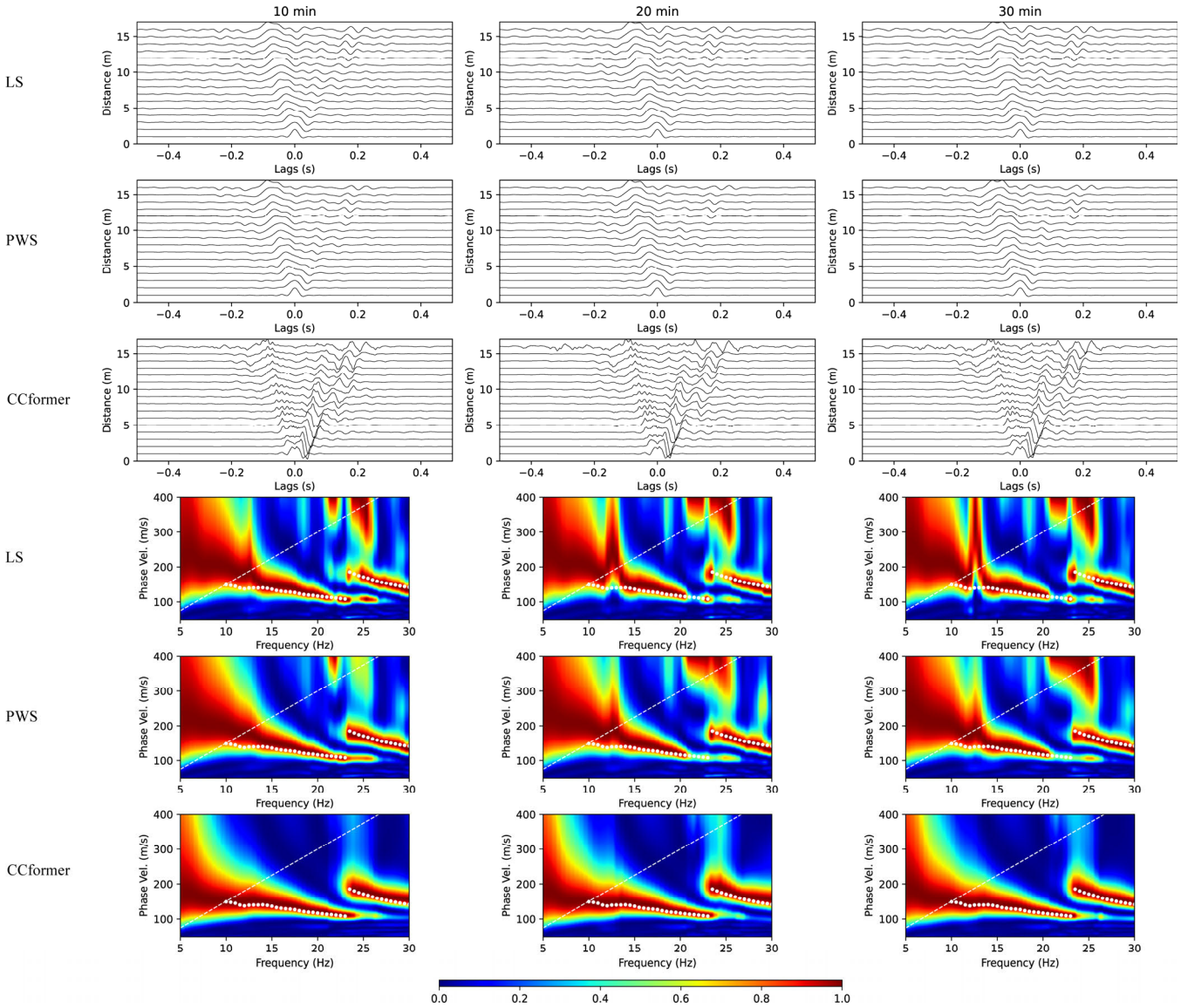


Fig. 15. Results for the field data of Line 17, with recording time of 10, 20, and 30 min. The first to third rows depict the cross-correlations produced by LS, PWS, and CCformer, respectively. The fourth to sixth rows represent the corresponding dispersion spectrum for LS, PWS, and CCformer results, respectively. The white dotted line in the dispersion spectrum indicates the spectral maxima on the dispersion spectrum of active data, and the white dashed line in the dispersion spectrum demonstrates the wavelength equal to the maximum array aperture. Note that all wiggle plots of the cross-correlations are on the same scale.

recordings. Fig. 15 depicts the stacked cross-correlations and corresponding dispersion spectrum for the first 10, 20, and 30 min of recordings produced by LS, PWS, and CCformer. The result indicates that although the cross-correlations of CCformer exhibit asymmetry due to insufficient recording time, the corresponding dispersion spectra for all cases present clear dispersive patterns that agree with the spectral maxima of active data. In contrast, the dispersion spectra of LS and PWS either exhibit energy gaps or overestimate phase velocities. The results of LS also suggest that the quality of cross-correlations retrieved from high-frequency ambient noise might not simply increase with recording time, as sources within nonstationary phase regions can be dominant at any time window. Nevertheless, the CCformer extracts coherent and effective information from cross-correlations of different time segments, always producing stacked cross-correlations with higher quality for situations with different recording times. Note that the recording time necessary to achieve high-quality cross-correlation also depends on the complexity of noise sources at the site [39]. Given the high efficiency of the CCformer in the inference stage (no more than 0.02 s for processing each two-station cross-correlation of our 24 h field data, consuming approximately 634 MB of GPU resources within our runtime environment), its combination with wireless transmission modules holds the potential of real-time estimation of data quality, thus avoiding unnecessary lengthy recordings. Moreover, considering the effectiveness of the CCformer, its extension to enhance body waves and process multicomponent ambient noise recordings could be a direction for future work.

## V. Conclusion

We have developed a novel deep learning-based algorithm to improve stacked high-frequency ambient noise surface wave cross-correlation for estimating surface wave dispersion. The proposed neural network takes advantage of the attention mechanism to pay attention to coherent and effective surface wave information between different time segments and with a design of integrating the process of stacking original individual cross-correlation by learning to produce a weight for each sample of each time segment. Tests on synthetic and field data all show that the proposed algorithm adequately enhances effectively coherent surface wave signals and attenuates spurious and incoherent noises, providing superior results compared to LS and PWS. Moreover, it presents good generalizability for datasets with different dispersions, source distributions, and acquisition parameters. It consequently provides a practical solution for automatically extracting effective surface wave signals from high-frequency ambient noise. Other possible developments might be extending the proposed improvement algorithm in enhancing body waves and processing multicomponent ambient noise recordings, which could extract more high-quality information about the subsurface from ambient noise.

## Acknowledgment

The authors thank Chunfeng Rao of SGEEG (Group) Company Ltd. and Ruiqing Shen of Tongji University for their kind help during the acquisition. They also thank editors and two anonymous reviewers for their constructive comments and suggestions, which greatly improved this article. The source code is available at https://github.com/shufangeo/CCformer

## References

[1] H. Okada, *The Microtremor Survey Method* (Geophysical Monographs Series), vol. 12. Tulsa, OK, USA: Society Exploration Geophysicists, 2003.

[2] M. N. Toksöz and R. T. Lacoss, "Microseisms: Mode structure and sources," *Science*, vol. 159, no. 3817, pp. 872–873, Feb. 1968, doi: 10.1126/science.159.3817.872.

[3] O. I. Lobkis and R. L. Weaver, "On the emergence of the green's function in the correlations of a diffuse field," *J. Acoust. Soc. Amer.*, vol. 110, no. 6, pp. 3011–3017, 2001, doi: 10.1016/s0041-624x(02)00156-7.

[4] N. M. Shapiro, M. Campillo, L. Stehly, and M. H. Ritzwoller, "High-resolution surface-wave tomography from ambient seismic noise," *Science*, vol. 307, no. 5715, pp. 1615–1618, Mar. 2005, doi: 10.1126/science.1108339.

[5] F. Li, M. Valero, Y. Cheng, and W. Song, "High-frequency time-lapse seismic spatial autocorrelation imaging shallow velocity variations," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 12, pp. 5385–5390, Dec. 2019, doi: 10.1109/JSTARS.2019.2954114.

[6] J. Shao, Y. Wang, and L. Chen, "Near-surface characterization using high-speed train seismic data recorded by a distributed acoustic sensing array," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5912911, doi: 10.1109/TGRS.2022.3153831.

[7] B. Mi and J. Xia, "Extraction of Rayleigh, love, and virtual refraction waves from 3C high-speed-train-induced vibrations for near-surface characterization," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5919710, doi: 10.1109/TGRS.2023.3318989.

[8] R. Snieder, "Extracting the green's function from the correlation of coda waves: A derivation based on stationary phase," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, no. 4, Apr. 2004, Art. no. 046610, doi: 10.1103/physreve.69.046610.

[9] R. Snieder and E. Larose, "Extracting Earth's elastic wave response from noise measurements," *Annu. Rev. Earth Planet. Sci.*, vol. 41, no. 1, pp. 183–206, May 2013, doi: 10.1146/annurev-earth-050212-123936.

[10] I. Gaudot, É. Beucler, A. Mocquet, M. Schimmel, and M. Le Feuvre, "Statistical redundancy of instantaneous phases: Theory and application to the seismic ambient wavefield," *Geophys. J. Int.*, vol. 204, no. 2, pp. 1159–1163, Dec. 2015, doi: 10.1093/gji/ggv501.

[11] H. Yao and R. D. Van Der Hilst, "Analysis of ambient noise energy distribution and phase velocity bias in ambient noise tomography, with application to SE Tibet," *Geophys. J. Int.*, vol. 179, no. 2, pp. 1113–1132, Nov. 2009, doi: 10.1111/j.1365-246X.2009.04329.x.

[12] M. Schimmel and H. Paulssen, "Noise reduction and detection of weak, coherent signals through phase-weighted stacks," *Geophys. J. Int.*, vol. 130, no. 2, pp. 497–505, Aug. 1997, doi: 10.1111/j.1365-246X.1997.tb05664.x.

[13] M. Schimmel, E. Stutzmann, and J. Gallart, "Using instantaneous phase coherence for signal extraction from ambient noise data at a local to a global scale," *Geophys. J. Int.*, vol. 184, no. 1, pp. 494–506, Jan. 2011, doi: 10.1111/j.1365-246X.2010.04861.x.

[14] J. Xie, Y. Yang, and Y. Luo, "Improving cross-correlations of ambient noise using an RMS-ratio selection stacking method," *Geophys. J. Int.*, vol. 222, no. 1, pp. 989–1002, Mar. 2020, doi: 10.1093/gji/ggaa232.

[15] B. Guan et al., "Improving data quality of three-component measurements of noise in urban environments using polarization analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5925610, doi: 10.1109/TGRS.2024.3435068.

[16] L. Moreau, L. Stehly, P. Boué, Y. Lu, E. Larose, and M. Campillo, "Improving ambient noise correlation functions with an SVD-based Wiener filter," *Geophys. J. Int.*, vol. 211, no. 1, pp. 418–426, Oct. 2017, doi: 10.1093/gji/ggx306.

[17] L. Ning, J. Xia, T. Dai, H. Zhang, Y. Liu, and Y. Hong, "Improving the quality of high-frequency surface waves retrieved from ultrashort traffic-induced noise based on eigenvalue selection," *Geophys. J. Int.*, vol. 235, no. 3, pp. 2020–2034, Sep. 2023, doi: 10.1093/gji/ggad343.

[18] Y. Liu, J. Xia, C. Xi, T. Dai, and L. Ning, "Improving the retrieval of high-frequency surface waves from ambient noise through multichannel-coherency-weighted stack," *Geophys. J. Int.*, vol. 227, no. 2, pp. 776–785, Aug. 2021, doi: 10.1093/gji/ggab253.

[19] L. Ning, J. Xia, T. Dai, Y. Liu, H. Zhang, and C. Xi, "High-frequency surface-wave imaging from traffic-induced noise by selecting in-line sources," *Surveys Geophysics*, vol. 43, no. 6, pp. 1873–1899, Aug. 2022, doi: 10.1007/s10712-022-09723-2.

[20] A. M. Baig, M. Campillo, and F. Brenguier, "Denoising seismic noise cross correlations," *J. Geophys. Research: Solid Earth*, vol. 114, no. B8, Aug. 2009, Art. no. B08310, doi: 10.1029/2008jb006085.

[21] G. Melo, A. Malcolm, D. Mikesell, and K. van Wijk, "Using SVD for improved interferometric green's function retrieval," *Geophys. J. Int.*, vol. 194, no. 3, pp. 1596–1612, Sep. 2013, doi: 10.1093/gji/ggt172.

[22] L. Viens and T. Iwata, "Improving the retrieval of offshore-onshore correlation functions with machine learning," *J. Geophys. Res., Solid Earth*, vol. 125, no. 8, Aug. 2020, Art. no. e2020JB019730, doi: 10.1029/2020jb019730.

[23] L. Viens and C. Van Houtte, "Denoising ambient seismic field correlation functions with convolutional autoencoders," *Geophys. J. Int.*, vol. 220, no. 3, pp. 1521–1535, Mar. 2020, doi: 10.1093/gji/ggz509.

[24] H. Sun and L. Demanet, "Beyond correlations: Deep learning for seismic interferometry," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 7, pp. 3385–3396, Jul. 2023, doi: 10.1109/TNNLS.2022.3172385.

[25] B. Zhao, L. Han, P. Zhang, and Y. Yin, "Noise reduction and encrypted reconstruction of passive source virtual shot records based on GMF-RS network," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5904410, doi: 10.1109/TGRS.2024.3357724.

[26] J. F. Lawrence, M. Denolle, K. J. Seats, and G. A. Prieto, "A numeric evaluation of attenuation from ambient noise correlation functions," *J. Geophys. Res., Solid Earth*, vol. 118, no. 12, pp. 6134–6145, Dec. 2013, doi: 10.1002/2012JB009513.

[27] A. Vaswani et al., "Attention is all you need," in *Proc. NIPS*, 2017, pp. 6000–6010.

[28] Z. H. Zhang et al., "Deep learning inversion of Rayleigh dispersion curves," *Chin. J. Geophys.*, vol. 65, no. 6, pp. 2244–2259, Jun. 2022, doi: 10.6038/cjg2022P0446.

[29] N. A. Haskell, "The dispersion of surface waves on multilayered media," *Bull. Seismol. Soc. Amer.*, vol. 43, no. 1, pp. 17–34, 1953, doi: 10.1785/BSSA0430010017.

[30] W. T. Thomson, "Transmission of elastic waves through a stratified solid medium," *J. Appl. Phys.*, vol. 21, no. 2, pp. 89–93, Feb. 1950, doi: 10.1063/1.1699629.

[31] J. W. Thorbecke and D. Draganov, "Finite-difference modeling experiments for seismic interferometry," *GEOPHYSICS*, vol. 76, no. 6, pp. 1–18, Nov. 2011, doi: 10.1190/geo2010-0039.1.

[32] G. D. Bensen et al., "Processing seismic ambient noise data to obtain reliable broad-band surface wave dispersion measurements," *Geophys. J. Int.*, vol. 169, no. 3, pp. 1239–1260, Jun. 2007, doi: 10.1111/j.1365-246x.2007.03374.x.

[33] N. Nakata, R. Snieder, T. Tsuji, K. Larner, and T. Matsuoka, "Shear wave imaging from traffic noise using seismic interferometry by cross-coherence," *Geophysics*, vol. 76, no. 6, pp. SA97–SA106, Nov. 2011, doi: 10.1190/geo2010-0188.1.

[34] F.-C. Lin, M. P. Moschetti, and M. H. Ritzwoller, "Surface wave tomography of the western United States from ambient seismic noise: Rayleigh and love wave phase velocity maps," *Geophys. J. Int.*, vol. 173, no. 1, pp. 281–298, Apr. 2008, doi: 10.1111/j.1365-246X.2008.03720.x.

[35] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5884–5888, doi: 10.1109/ICASSP.2018.8462506.

[36] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent.*, Dec. 2018, pp. 1–16.

[37] L. V. Socco and C. Strobbia, "Surface-wave method for near-surface characterization: A tutorial," *Near Surf. Geophys.*, vol. 2, no. 4, pp. 165–185, Nov. 2004, doi: 10.3997/1873-0604.2004015.

[38] S. Hu, Y. Zhao, L. V. Socco, and S. Ge, "Retrieving 2-D laterally varying structures from multistation surface wave dispersion curves using multiscale window analysis," *Geophys. J. Int.*, vol. 227, no. 2, pp. 1418–1438, Aug. 2021, doi: 10.1093/gji/ggab282.

[39] L. Ehsaninezhad, C. Wollin, V. Rodríguez Tribaldos, B. Schwarz, and C. M. Krawczyk, "Urban subsurface exploration improved by denoising of virtual shot gathers from distributed acoustic sensing ambient noise," *Geophys. J. Int.*, vol. 237, no. 3, pp. 1751–1764, Apr. 2024, doi: 10.1093/gji/ggae134.

**Shufan Hu** (Member, IEEE) received the bachelor's degree in exploration technology and engineering from East China University of Technology, Nanchang, China, in 2014, and the master's and Ph.D. degrees in geophysics from Tongji University, Shanghai, China, in 2017 and 2021, respectively.

He was a Visiting Ph.D. Student with the Politecnico di Torino, Turin, Italy, from 2018 to 2019. He is currently a Lecturer with the Department of Computer Science and Technology, Nanchang University, Nanchang. His research interests include geophysical signal processing, inverse problems, and machine learning.

**Huilin Zhou** received the Ph.D. degree in space physics from Wuhan University, Wuhan, China, in 2006.

In 2011, he was a Visiting Scholar with the State Key Laboratory of Radar Signal Processing, Xidian University, Xi'an, China. He is currently a Professor with the Multi-Functional Optical/RF Sensor and AI (FORSEAI) Laboratory and the School of Information Engineering, Nanchang University, Nanchang, China. His research interests include radar systems, geophysical signal processing, and radar imaging.

**Laura Valentina Socco** received the Ph.D. degree in geoenvironmental engineering from Politecnico di Torino, Turin, Italy, in 1996.

She is currently a Full Professor of applied geophysics with Delft University of Technology, Delft, The Netherlands, and also with Politecnico di Torino. Her research focuses on developing geophysical methods based on seismic surface wave and data integration in applications ranging from, seismic hazard and engineering, to hydrocarbon exploration, environment, and cultural heritage.

Dr. Socco is currently the EAGE President. In 2013, she was chosen as an Honorary Lecturer by the SEG. In 2014, she received the Conrad Schlumberger Award (EAGE). In 2019, she received the Outstanding Educator Award from SEG. She was the Editor-in-Chief of Geophysics from 2017 to 2019.

**Yonghui Zhao** received the bachelor's degree in exploration geophysics from Changchun College of Geology, Changchun, China, in 1996, and the master's and Ph.D. degrees in solid geophysics from Tongji University, Shanghai, China, in 1999 and 2001, respectively.

He held a post-doctoral position at the College of Civil Engineering, Tongji University, from 2001 to 2003. In 2010, he was a Visiting Scholar with the Electro Science Laboratory, The Ohio State University, Columbus, OH, USA. He is currently an Associate Professor with the School of Ocean and Earth Science, Tongji University. His research interests include ground penetrating radar, engineering and environmental geophysics, and integrated geophysics.