Jin *et al.*
EURASIP Journal on Advances in Signal Processing (2022) 2022:107
https://doi.org/10.1186/s13634-022-00941-9

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

Semi-supervised underwater acoustic source localization based on residual convolutional autoencoder

Pian Jin[1], Biao Wang[1,3*], Lebo Li[2], Peng Chao[1] and Fangtong Xie[1]

*Correspondence: wangbiao@just.edu.cn

[1] Jiangsu University of Science and Technology, Zhenjiang, China
[2] Hangzhou Applied Acoustics Research Institute, Hangzhou, China
[3] Science and Technology on Underwater Vehicle Technology Laboratory, Harbin Engineering University, Harbin, China

Abstract

Passive localization of underwater targets was a thorny problem in underwater acoustics. For traditional model-driven passive localization methods, the main challenges are the inevitable environmental mismatch and the presence of interference and noise everywhere. In recent years, data-driven machine learning approaches have opened up new possibilities for passive localization of underwater acoustics. However, the acquisition and processing of underwater acoustics data are more restricted than other scenarios, and the lack of data is one of the most enormous difficulties in the application of machine learning to underwater acoustics. To take full advantage of the relatively easy accessed unlabeled data, this paper proposes a framework for underwater acoustic source localization based on a two-step semi-supervised learning classification model. The first step is trained in unsupervised mode with the whole available dataset (labeled and unlabeled dataset), and it consists of a convolutional autoencoder (CAE) for feature extraction and self-attention (RA) mechanism for picking more useful features by applying constraints on the CAE. The second step is trained in supervised mode with the labeled dataset, and it consists of a multilayer perceptron connected to an encoder from the first step and is used to perform the source location task. The proposed framework is validated on uniform vertical line array data of SWellEx-96 event S5. Compared with the supervised model and the model without the RA, the proposed framework maintains good localization performance with the reduced labeled dataset, and the proposed framework is more robust when the training dataset and the test dataset of the second step are distributed differently, which is called "data mismatch."

Keywords: Underwater acoustic source localization, Semi-supervised learning, Convolutional autoencoder, Self-attentive mechanism, Data mismatch

1 Introduction

Passive localization of underwater targets has been a complex problem in the underwater acoustics field. Unlike free-field environments, underwater acoustic channels are typically characterized by multipath and spatiotemporal variability, which cannot be ignored in both source localization and communication. Multipath structure allows us to make more accurate estimates of source location, including source depth and range. The spatiotemporal variability makes it difficult to master the precise channel parameters

which also called waveguide parameters. To achieve effective localization of acoustic source in an ocean waveguide, an accurate underwater acoustic propagation model and prior waveguide parameters are inseparable from the traditional model-driven approach. Matched-field processing (MFP) [1–5] has been one of the primary methods for passive localization of underwater targets in the last three decades. In the natural ocean environment, the uncertainty of environmental information seriously affects the performance of MFP [6, 7], which is called environmental mismatch in the underwater acoustic field. In response to the mismatch problem, a series of improved MFP methods have emerged. For example, Focalization MFP [8, 9], proposes that the sound source localization and the environment parameters should be searched simultaneously.

In recent years, data-driven machine learning approaches have contributed to the development of acoustic signal processing [10–12]. Machine learning approaches can be considered offline training and online prediction strategy. A large amount of intensive computation is concentrated in the training phase of the model, and the trained model performs the lightweight analysis in the prediction phase, so real-time processing of data can be achieved more easily; where deep learning methods use deeper network structures and have better feature extraction capabilities compared to shallow networks [13]. The training of deep neural networks relies on "big data, " it can be said that the deep neural network and big data together help the model to be closer to the measured data distribution in a statistical sense and thus obtain better prediction performance.

In the underwater acoustics field, machine learning has also been applied in various aspects, such as detection/classification and localization of underwater targets [14–18]. It has also been used for seabed classification and ocean environmental information extraction [11, 19–21], and has produced rich results. In addition, a large number of research progresses related to machine learning methods have also appeared in the field of underwater acoustic communication [22–28]. At the same time, there are many challenges in using machine learning for underwater acoustic signal processing; mainly, the process of acquiring datasets has many limitations, especially the labeled datasets, which makes it difficult to form "big data" conditions. A trend has emerged: using sound field model simulation labeled data instead of natural measurement labeled data [29] to train supervised learning models. For example, Haiqiang Niu et al. [30] trained a deep residual network by simulated data and tested it on measured data, achieving better results than the focalization MFP, the cost of acquiring data for this method is low, but to compensate for the distribution differences between simulated and measured data (because environmental parameters used to create the simulated dataset always not be able to explain the measured data), it is necessary to simulate a large number of training datasets under different environmental parameters for improving the generalization ability of the network, and the training cost is high; for the problem of data distribution differences, some scholars have applied migration learning [31, 32] to passive localization of underwater sources and the strategy of adding fine-tuning to pre-training is proposed, using a small amount of measured data to fine-tune the network model trained with simulation data.

Inspired by the above articles, we propose a method based on the Semi-supervised learning (SSL) model, and we regard source localization as a classification problem. In the first step, a CAE added the residual self-attention mechanism (RA-CAE) is used to perform the feature extraction for the whole dataset by unsupervised learning. The

second step uses the encoder trained in the first step to extract features from the labeled data; then, the features are classified by a 4-layer multilayer perceptron (MLP) to perform source localization task. Together, the two steps constitute a semi-supervised learning framework (RA-CAE-SSL) for source localization.The performance of our method is evaluated using VLA reception data from the SWellEx-96 experimental S5 event [33].

The structure of the paper is as follows: Sect. 2 introduces the proposed two-step framework RA-CAE-SSL and theories related to Self-attention mechanism and CAE, as well as the performance evaluation indexes; Sect. 3 shows the data pre-processing process and analyze the Swellex-96 VLA data by traditional conventional signal processing methods, including MFP; in Sect. 4, we firstly propose two ways to divide the dataset; then, we give the corresponding localization results and discuss the performance of our proposed framework by comparing it with the control groups; and Sect. 5 shows the conclusions as well as future work.

2 Proposed method based on the semi-supervised learning model

The proposed method is performed through two steps: the first step is training the feature extraction part by unsupervised learning which is consisted of convolution autoencoder and self-attention model, and the second step is train classify part by supervised learning which is consisted of feature extraction with fixed weights and multilayer perceptron. In this section, we firstly introduce the convolution autoencoder and self-attention model that we used, then introduce the two-step framework.

2.1 Convolutional autoencoder

Convolutional autoencoder (CAE) [34, 35] is a kind of artificial neural network used in unsupervised learning, which uses convolution kernel for feature extraction. It can reduce the number of network parameters through weight sharing and local awareness features, while improving the model's ability to extract local features from the data.

The working principle of the CAE is shown in Fig. 1, the convolutional transformation process from feature mapping input to output is called convolutional encoder, and the output value is reconstructed by transposed convolution operation, called convolution decoder, where T represents the convolutional encode operation and T' represents the convolutional decode operation. The input feature matrix is $x \in R^{n \times Q \times Q}$. It contains n feature matrices, and the size of each feature matrix is $Q \times Q$.

2.2 Self-attention mechanism

The capabilities we expect from CAE are not simply copying input to output, and we would like to add some constraints to the CAE, so that the model will be forced to consider which parts of the input are much critical and need to be copied firstly. For example: undercomplete autoencoder, regular autoencoder, denoising autoencoder, etc. For the network to extract better features about the location information of the source, this paper uses the self-attention (SA) mechanism to impose constraints on the CAE. The SA was first applied to natural language processing [36, 37]. The traditional convolution operation extracts features based on the weights of the convolution filter over a local perceptual field using an aggregation function, and these weights are shared throughout

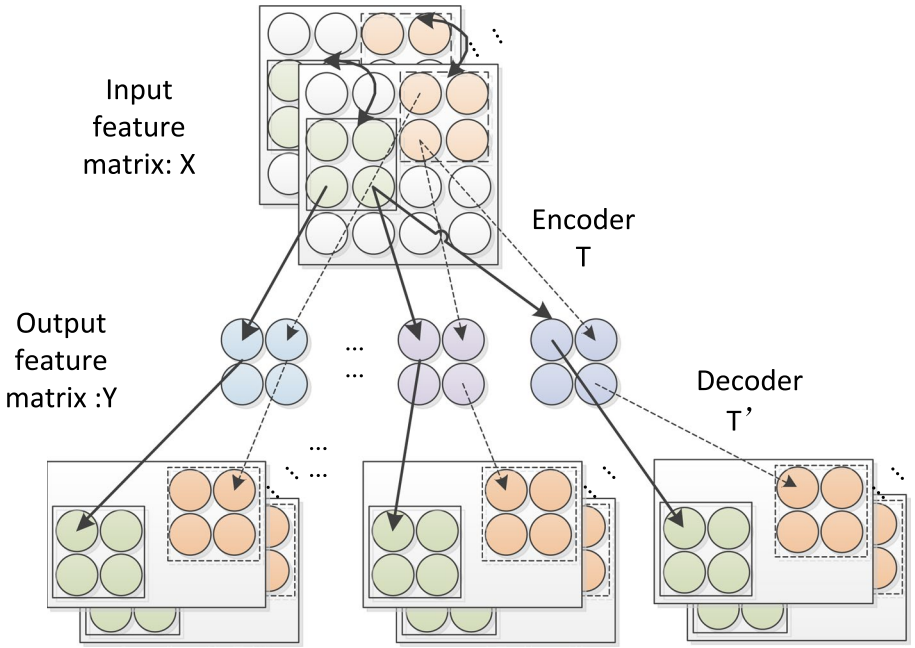

Fig. 1 Schematic diagram of a convolutional autoencoder

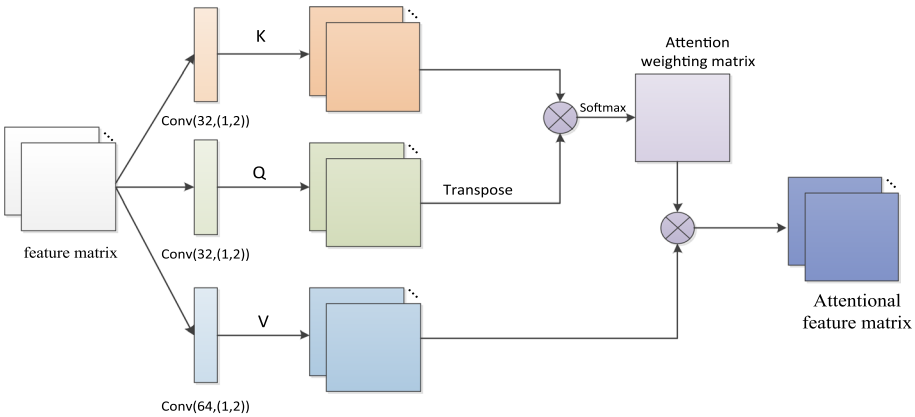

Fig. 2 Structure of self-attention model

the feature matrices. In contrast, the Self-Attention (SA) module uses a weighted average operation based on the input features to dynamically calculate the attention weights by correlation operations on the similarity function between features [38, 39].

Considering the different and complementary nature of convolutional operations and SA, there exists the potential to benefit from both paradigms through integration, so this paper combines the CAE with the self-attentive mechanism to propose the CAE with the Residual self-attention mechanism module (RA-CAE), introducing the residual module can make the training process more efficient by the ability to transform through identity.

The model structure of SA is shown in Fig. 2, where the output feature tensor $\boldsymbol{X} \in \boldsymbol{R}^{C_{in} \times W \times H}$ of the convolutional layer in the CAE is used as the input of the layer,

where H and W denote the dimensions of the tensor, let $x_{ij} \in \boldsymbol{R}^{C_{in}}$ denote the elements of the input tensor, let $\boldsymbol{Y} \in \boldsymbol{R}^{C_{\text{out}} \times W \times H}$ denote the output feature matrix, and let $y_{ij} \in \boldsymbol{R}^{C_{out}}$ denote the components of the output tensor. Let:

$$\begin{aligned} \boldsymbol{q}_{ij} &= \boldsymbol{W}^{q} x_{ij} \\ \boldsymbol{k}_{ij} &= \boldsymbol{W}^{k} x_{ij} \\ \boldsymbol{v}_{ij} &= \boldsymbol{W}^{v} x_{ij} \end{aligned} \tag{1}$$

Then, the output of the SA can be expressed as:

$$\begin{aligned} y_{ij} &= \sum_{a,b \in \mathcal{N}_k(i,j)} \text{Attention}\left(q_{ij}, k_{ab}\right) v_{ab} \\ &= \sum_{a,b \in \mathcal{N}_k(i,j)} \text{softmax}\left(\frac{\left(W^{q} x_{ij}\right)^{T} \left(W^{k} x_{ab}\right)}{\sqrt{d}} \right) W^{v} x_{ab} \end{aligned} \tag{2}$$

where W^{q}, W^{k}, W^{v} denotes the weight matrix, $\mathcal{N}_k(i,j)$ denotes a local region centered at (i,j) with spatial extent k, S_{ij} denotes the attention weight of features in the region $\mathcal{N}_k(i,j)$, and d denotes the feature dimension of $W^{q} x_{ij}$.

In this paper, the SA projects the feature matrix output from the autoencoder Conv2 as Q, K, V using a convolution kernel of 1×2. After that the attention weights are computed and matrix aggregation is performed to extract the local features of the classified objects.

2.3 Proposed model framework: RA-CAE-SSL

In underwater acoustics, dataset acquisition is limited, especially reliable labeled dataset, which are difficult to form "big data" conditions, and this also limits the application of deep supervised learning models to underwater acoustic source localization. In this paper, we propose a two-step semi-supervised learning framework under the assumption that labeled dataset are insufficient and unlabeled dataset are relatively abundant. The specific steps are as follows:

Step 1: Training the RA-CAE model

The first step performs unsupervised learning on RA-CAE model to achieve coverage of the entire dataset (unlabeled data and labeled data).

The structure of RA-CAE is shown in Fig. 3a. The encoder consists of three convolutional modules can project the input data into the hidden space; The decoder has a symmetrical structure with the encoder and is dedicated to reconstructing the input data from the hidden space; The residual block with self-attention is placed between the encoder and the decoder and serves to place attention on the features that are more important. The whole dataset (including unlabeled part and labeled part) will be used in this step as training dataset since it doesn't need additional category information, and the loss function of mean square error (MSE).

Step 2: Training RA-CAE-SSL model for source localization

The second step performs supervised learning on the RA-CAE-SSL model.

After completing the training of the RA-CAE model, taking out part of the structure in Fig. 3a and freezing the parameters as a feature extraction network, which is connected

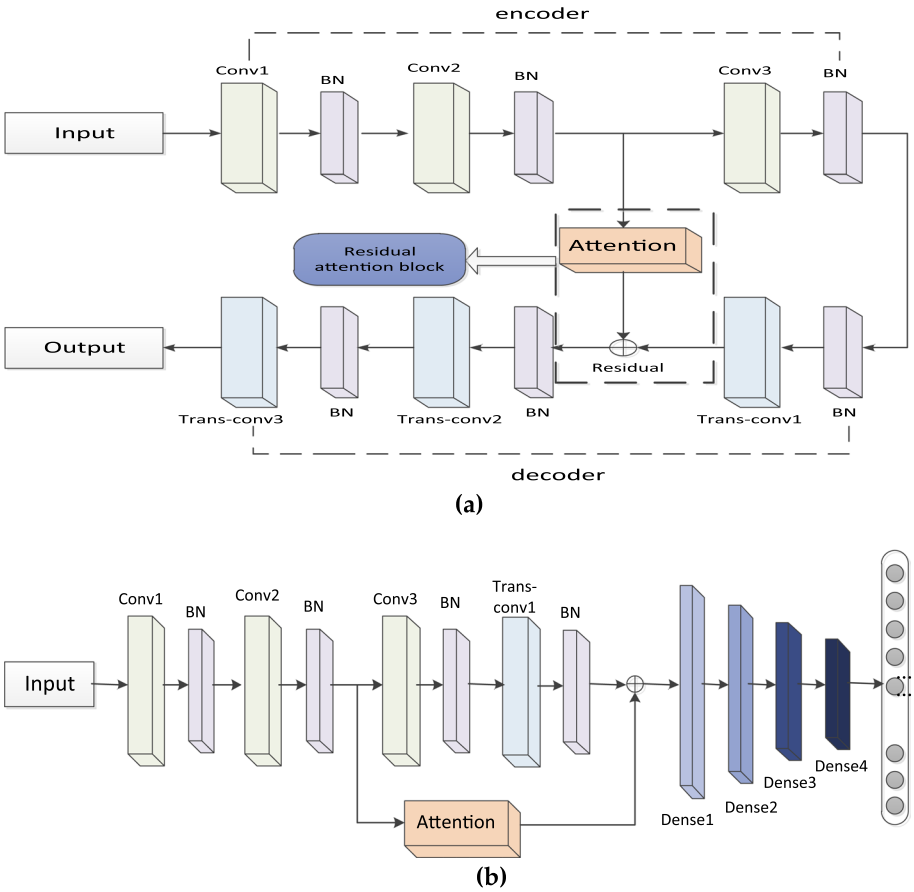

Fig. 3 Underwater acoustic source localization network framework design: **a** the structure of RA-CAE model; **b** the structure of RA-CAE-SSL model

with a 4-layer MLP classification network to form the RA-CAE-SSL model, whose construction is shown in Fig. 3b The labeled dataset is first passed through a trained feature extraction network. Then, the extracted features are fed into the MLP for classification learning to achieve the source localization task with a loss function of Cross-Entropy Loss Function (CELF).

2.4 Performance metrics

The commonly used evaluation metrics in traditional sound source localization are Mean Absolute Error (MAE) and Probability of credible localization (P_{CL}), and the total number of samples is S. The actual distance corresponding to the i th sample is y_i, and the predicted value is $f(x_i)$.

The MAE is calculated by the following formula:

$$\mathrm{MAE} = \frac{1}{S}\sum_{i=1}^{S}\left|y_i - f(x_i)\right| \tag{3}$$

P_{CL} specifies an error limit, and considers all samples falling within the error limit as correctly predicted samples, and calculates the localization accuracy from this. For example, at the 5% error limit, the localization accuracy $P_{CL-5\%}$ is calculated as follows:

$$P_{CL-5\%} = \frac{\sum_{i=1}^{S} \eta(i)}{S} \tag{4}$$

where

$$\eta(i) = \begin{cases} 1, & \dfrac{|y_i - f(x_i)|}{y_i} \times 100\% \leq 5\% \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

A smaller MAE value indicates better positioning performance, and a larger P_{CL} value indicates better positioning performance.

2.5 Difference from MFP

(1) The execution strategies and efficiency of the algorithms are different: Machine learning methods can be thought of as an offline training, online prediction strategy.
(2) The cost function used for localization is different: machine learning methods mostly use cost functions such as minimum mean square error or minimum cross-entropy training. The matching field processing mostly adopts the method of correlation processing.

3 Data pre-processing and SWellEX-96 data analysis

3.1 Data pre-processing

The sound pressure field under the ocean waveguide acoustic propagation model can be modeled as:

$$p(f) = S(f) \cdot g(f, r_s, r_m) + \varepsilon \tag{6}$$

where $p(f)$ is the complex acoustic pressure at the receiving array element, which can be obtained by discrete Fourier transform (DFT) of the original acoustic pressure data received by the array element, $S(f)$ is the source term, $g(f, r_s, r_m)$ is the Green's function to describe the channel response between the source position r_s and the receiving array element position r_m, and ε is the ocean noise.

In the traditional underwater acoustic source localization methods, the sampling covariance matrix (SCM) of the receiving array is one of the commonly extracted features, which contains the position information of the source and the marine environment parameter information. In this paper, the SCM of the VLA is used as the feature input of the network.

Suppose that Q vertical array elements receive the complex sound pressure:

$$\mathbf{P}_\theta(f) = \left[p_1(f), p_2(f), ..., p_Q(f)\right]^T \tag{7}$$

where θ is the location.

To perform the normalization operation:

$$\tilde{\mathbf{P}}_\theta(f) = \frac{\mathbf{P}_\theta(f)}{\sqrt{\sum_{q=1}^{Q} \left|p_q(f)\right|^2}} = \frac{\mathbf{P}_\theta(f)}{\left\|\mathbf{P}_\theta(f)\right\|_2} \tag{8}$$

The SCM is calculated based on the average of the L snapshot data to obtain:

$$SCM_\theta(f) = \frac{1}{L}\sum_{l=1}^{L} \tilde{\mathbf{P}}_{l,\theta}(f)\tilde{\mathbf{P}}_{l,\theta}^H(f) \tag{9}$$

Taking the real and imaginary parts of the SCM matrix to obtain two $Q \times Q$ dimensional real matrices $SCM1$ and $SCM2$, and the real matrix is scaled to the interval (0,1) by the min–max scaling method:

$$SCM1 = \frac{SCM1 - SCM1_{\min}}{SCM1_{\max} - SCM1_{\min}} \tag{10}$$

$$SCM2 = \frac{SCM2 - SCM2_{\min}}{SCM2_{\max} - SCM2_{\min}} \tag{11}$$

The input to the semi-supervised network is a normalized covariance matrix of dimension $Q \times Q \times 2N$, where N is the number of frequency points.

3.2 Data label processing

Assuming that the source distance range is $(r_{\min}, \gamma_{\max}]$, using equal-width split-box discretization, the source distance is divided into K categories, that is:

$$\Delta r = \frac{r_{\max} - r_{\min}}{K} \tag{12}$$

Then, the generation of the label of the ith sample becomes:

$$\text{label}_i = \left\lceil \frac{r_i - r_{\min}}{\Delta r} \right\rceil \tag{13}$$

where Δr is the distance interval corresponding to the category, r_i is the distance between the i th sample and the receiving array, and $\lceil\ \rceil$ denotes the upward rounding function.

The actual distance of the sample belonging to the category label_i is processed by One-Hot Encoding and mapped to a $1 \times K$ binary label vector, and the value of K in this paper is taken as 100.

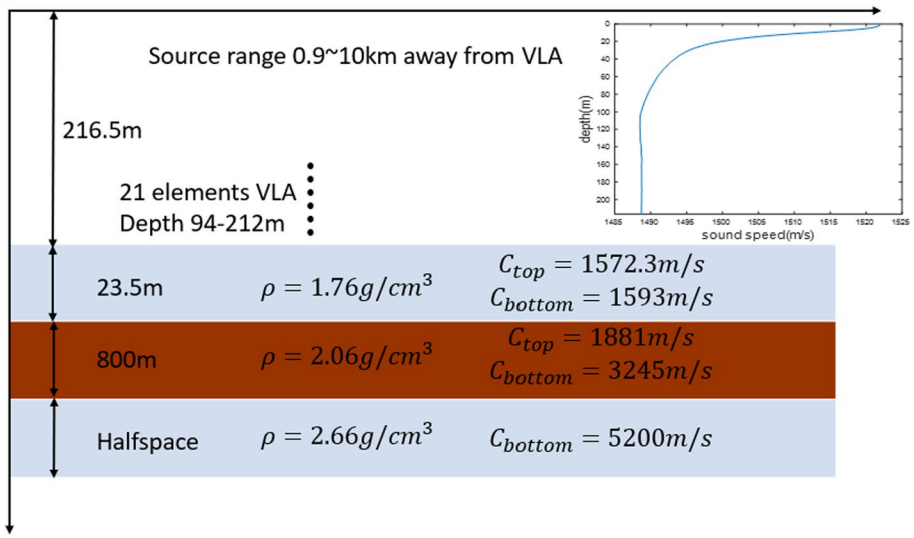

Fig. 4 SWellEX-96 experimental environment parameters

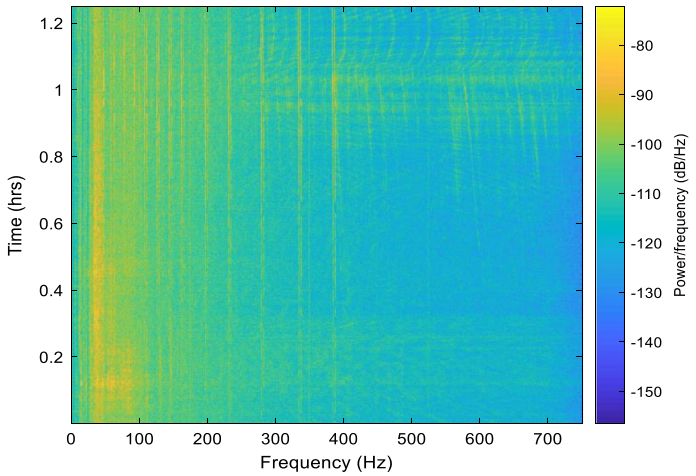

Fig. 5 The frequency spectrum of the signal received by array element No. 1

3.3 SWellEX-96 data analysis

The data of the VLA in event S5 of the SWellEX-96 experiment were used in this paper. The SWellEX-96 experiment was conducted at Point Loma, near San Diego, CA, from May 10 to May 18, 1996, and the environmental parameters of the sea area are shown in Fig. 4. The VLA with 21 hydrophones were placed at equal intervals in the sea depth range of 94.125 m to 212.25 m, with an array aperture of 118.125 m and a sampling frequency of 1500 Hz. The experimental vessel sailed from south to north, and the towed acoustic source emitted CW signals at {109, 127, 145, 163, 198, 232, 280, 335, 385} Hz at a source depth of 9 m, the VLA recorded the full 75 min event.

The time–frequency diagram of the received signal of the first hydrophone of the VLA is shown in Fig. 5. The signal-to-noise ratio of the actual data is estimated according to Eq. (14), which shows that in the first half of the voyage, the spectral value at the CW signal frequency is lower than the second half due to the long distance of the sound source

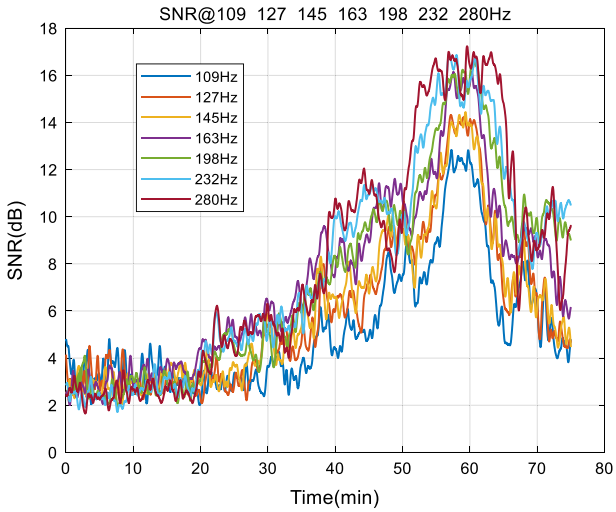

Fig. 6 The signal-to-noise ratio of the signal at each frequency point of array element No. 1 with time

from the array. From Fig. 6, we can know that the signal-to-noise ratio of the received signal increases when the sound source is close to the VLA.

$$\mathrm{SNR} \approx 10\log_{10}\left(\frac{Tr(\mathbf{C}_r)}{Tr(\mathbf{C}_n)} - 1\right) \tag{14}$$

where $\mathbf{C}_r$ is the signal covariance matrix, $\mathbf{C}_n$ is the noise covariance matrix.

MFP results

To make a preliminary analysis of the quality of the VLA array data and compare it with the proposed method. We processed the VLA data with MFP firstly. MFP is a generalized beamforming method which uses the spatial complexities of acoustic fields in an ocean waveguide to localize sources. The Bartlett MFP formula is as follows:

$$\mathbf{B}(\hat{\theta}) = \sum_{i=1}^{N_f} \mathbf{P}_{\hat{\theta}}^{H}(f_i)\mathbf{SCM}_{\theta}(f_i)\mathbf{P}_{\hat{\theta}}(f_i) \tag{15}$$

where θ is the location parameter, $\mathbf{P}(\hat{\theta})$ is the Steering vector, $\mathbf{B}(\hat{\theta})$ is the output of the beamforming.

The prior information required by MFP includes array parameters and waveguide parameters such as sound speed profile, depth of sea and sedimentary layer characteristics While the prior information required by the proposed method is a large number of datasets with different range. We can think of cost function of MFP as the distance between two vectors in Euclidean space. The MFP results are shown in Fig. 7, with the 10% error limit in the shaded part. The copy field is obtained from the Kraken model simulation, and the sound field model environment parameters are referred to Fig. 5. It can be seen that the matching field processing results are not satisfactory, and the narrowband matching field processing significantly degrades the prediction performance with a large number of discrete points when the sound source distance is greater than 4 km. The reduced signal-to-noise ratio is one of the reasons. The

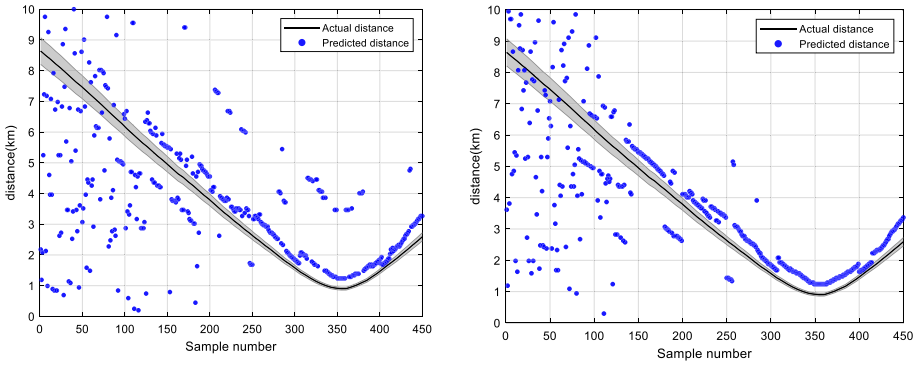

Fig. 7 MFP Processing results: **a** narrowband {163} Hz matching field results; **b** {109, 127, 145, 163, 198, 232, 280} Hz broadband matching field results

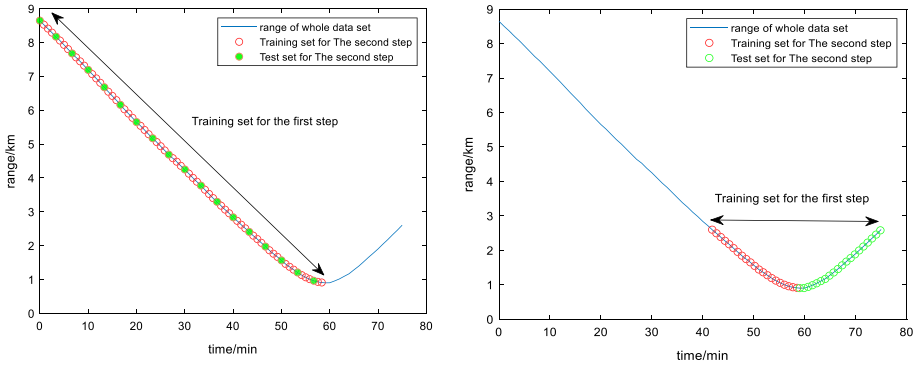

Fig. 8 Dataset division method: **a** The first type of dataset division; **b** the second type of dataset division

broadband matching field processing superimposed the ambiguity function of depth and range at each frequency point, which had the effect of enhancing the main lobe and suppressing the side lobe, and the anti-noise ability was more substantial than the narrowband matching field. Although the broadband matched field results can see the trend of sound source motion, there is a certain gap with the actual motion trajectory, and there are a small number of outlier points. This is mainly due to the mismatch of environmental parameters, especially the mismatch caused by the uncertainty of sea depth and the bottom parameters in the experimental sea area, beyond that the array location mismatch caused by fluctuation of water is also an important reason.

4 Source localization results and discussion

4.1 Datasets division and control group setting

4.1.1 Datasets division

Whether the data of training data and the test data satisfy the same distribution is an essential factor affecting the prediction performance of the model, to verify the implemen-tation of the semi-supervised framework proposed in this paper in the different cases, the data set is divided as follows, corresponding to the cases of the same

Table 1 Parameters of the network model for each part of the semi-supervised learning underwater acoustic source localization framework

	Block	Output channel	Kernel size	Stride
Encoder	Conv1	64	3 × 3	1
	Conv2	128	3 × 3	2
	Conv3	256	3 × 3	2
Decoder	Transposed-Conv1	256	3 × 3	2
	Transposed-Conv2	128	2 × 2	2
	Transposed-Conv3	64	4 × 4	1
MLP	Dense_1	4096	-	1
	Dense_2	2048	-	1
	Dense_3	1024	-	1
	Dense_4	100	-	1

distribution and different distributions, respectively, the data sets are divided in the way as follows:

Division 1: the data collected by VLA from 0 to 60 min were preprocessed to obtain 3540 samples, and they were used as the training sets for step 1; in the Step 2, we select two fractions from the whole sample set without repetition as the training set of the second step and the test set of the second step, every fraction selected should be uniformly distributed over the entire navigation path (Fig. 8a). Since the two fractions were selected from the same path, we think they approximately satisfy the same data distribution which is defined as "matched" case.

Division 2: the data collected by VLA from 45 to 75 min were preprocessed to obtain 2487 samples, and they were used as the training sets for step 1; in the Step 2, as Fig. 8b shown, the training sets for the step 2 is selected from the left side and the testing sets for the step 2 is selected from the right side. Since the two sets were selected from different path, we think they do not satisfy the same data distribution which is defined as "mismatched" case.

4.1.2 Control group setting

In order to comprehensively evaluate the performance of the proposed framework (RA-CAE-SSL) in underwater acoustic source localization, three control groups are proposed in this section.

Control group I (CAE-SSL): A semi-supervised learning approach is used to train a network model that only lacks the residual self-attention mechanism module compared to the RA-CAE-SSL;

Control group II (RA-SL): A supervised learning approach is used to train a network with the same structure as the RA-CAE-SSL;

Control group III (CNN): A supervised learning approach is used to train a network with the same structure as the CAE-SSL.

Table 2 Localization performance of narrowband dataset

	Percentage of label data	**75%**		**37.5%**		**15%**	
Frequency points	**Positioning network model**	**MAE (km)**	$P_{CL-5\%}$**(%)**	**MAE (km)**	$P_{CL-5\%}$**(%)**	**MAE (km)**	$P_{CL-5\%}$**(%)**
109	RA-CAE-SSL	0.9098	62	1.2392	55	1.5352	44.5
	CAE-SSL	0.9072	63.5	1.4963	50.5	1.7699	43.5
	RA-SL	0.8825	65	1.4024	51.5	1.8392	41.5
	CNN	0.7986	61.5	1.3297	46	1.9254	39.5
127	RA-CAE-SSL	0.6581	75.50	0.7100	72	0.9487	64
	CAE-SSL	0.6719	72.5	0.6489	72	1.2943	55.5
	RA-SL	0.7328	73	0.8213	67	1.1592	57.5
	CNN	0.7812	68	0.7570	66.50	1.2157	54.50
145	RA-CAE-SSL	0.3606	86	0.5211	77	0.6323	73
	CAE-SSL	0.4503	83	0.5954	76	0.6746	71
	RA-SL	0.4761	81	0.6873	73.5	0.8090	72
	CNN	0.4437	81.5	0.6897	72	0.7254	69
163	RA-CAE-SSL	0.1201	93.5	0.2259	89	0.4083	81
	CAE-SSL	0.1390	93	0.2679	89.5	0.5134	79
	RA-SL	0.1570	92	0.3891	86	0.5954	79.5
	CNN	0.1255	91.5	0.3094	86	0.7659	68
198	RA-CAE-SSL	0.1470	95.5	0.2802	86.5	0.4557	82
	CAE-SSL	0.1155	95.5	0.2090	90	0.4372	82
	RA-SL	0.1489	94	0.2428	88	0.5449	80
	CNN	0.1205	94	0.2413	87.5	0.5700	78

Table 3 Localization performance of broadband dataset

Frequency points	**Percentage of label data**	**75%**		**37.5%**		**15%**	
	Positioning network model	**MAE (km)**	$P_{CL-5\%}$**(%)**	**MAE (km)**	$P_{CL-5\%}$**(%)**	**MAE (km)**	$P_{CL-5\%}$**(%)**
①	RA-CAE-SSL	0.3964	82	0.3891	82.5	0.7024	70.5
	CAE-SSL	0.4657	80	0.6604	73.5	0.8721	64.5
	RA-SL	0.4426	81	0.7089	71	0.9729	66.5
	CNN	0.5115	79	0.7859	65	0.6600	65.5
②	RA-CAE-SSL	0.0724	96.5	0.2017	88.5	0.4961	78.5
	CAE-SSL	0.1474	93.5	0.2917	87	0.6400	77
	RA-SL	0.2232	90.5	0.4707	83.5	0.8701	71
	CNN	0.3506	81.5	0.4183	82.5	0.7301	65.5
③	RA-CAE-SSL	0.0100	100	0.0250	98.5	0.1589	91.5
	CAE-SSL	0.0439	98	0.0839	96.5	0.3671	89
	RA-SL	0.0535	98.5	0.0543	97	0.6296	82
	CNN	0.0252	98.5	0.2213	91	.0.5442	82
④	RA-CAE-SSL	0.0077	100	0.0115	99.5	0.0219	99
	CAE-SSL	0.0350	99	0.1135	94	0.1435	92
	RA-SL	0.0177	99.5	0.0466	97	0.1455	94
	CNN	0.0381	98.5	0.0562	96.5	0.2309	88.5

Where①: {109, 127}; ②: {109, 127, 145}; ③: {109, 127, 145, 163, 198}; ④: {109, 127145, 163, 198, 232, 280}.

4.2 Source localization results for the first dataset division method

To validate the performance of the proposed semi-supervised framework when the number of labeled data is reduced, we selected 75%, 37.5% and 15% of the entire labeled data as the training set for step 2, respectively. Meanwhile, in order to show the contribution of dataset bandwidth to localization performance, we conduct experiments using narrowband dataset consisting of single frequency and broadband dataset consisting of multiple frequencies, respectively, and the localization performance is shown as follows (Table 1).

Tables 2 and 3 give the localization performance under different model, different percentages of labeled dataset and different bandwidth of frequency.

We give the analysis of the results as follows:

1. Comparing the localization results of RA-CAE-SSL model and RA-SL model, it is found that the localization performance of the semi-supervised learning model out-performs the supervised learning in most cases, especially after the number of labeled data is reduced, the localization performance of the supervised learning model decreases, and the advantage of the semi-supervised learning model is more obvious, so we can say semi-supervised learning strategy is more suitable for underwater acoustic source localization when labeled data are not enough but unlabeled data are relatively abundant;
2. Comparing the localization results of RA-CAE-SSL model and CAE-SSL model, it is found that the introduction of the residual self-attention mechanism module can effectively improve the localization performance of the semi-supervised learning model, and it can be more useful when training data are insufficient;
3. Comparing the test results of broadband dataset and narrowband dataset, it is found that the localization performance of broadband datasets is better than that of narrowband datasets, and the more frequency points the samples contain, the better the localization performance. This is because the broadband samples carry more location information, which effectively reducing the uncertainty. This also appears in the matching field processing results;
4. Comparing the localization results of narrowband dataset at different frequency points, it is found that the localization performance of high-frequency dataset is better than that of low-frequency datasets. We speculate that it is due to there is more significant variation between different elements within the covariance matrix of high-frequency dataset, which is more conducive to feature extraction. This is consistent with the rule in the conventional beam formation, that the higher the frequency of the signal source, the smaller the main lobe is.
5. We find the performance of 198Hz is clearly better than other single frequency. From the perspective of normal modes, the higher the frequency of the source means the more normal wave modes are excited and therefore more information about the location of the source is contained in the signal. So it is understandable.

To give a more visual comparison, we chose sample localization result of dataset whose frequency is 163 Hz and {109,127,145,163,198,232,280} Hz, then plot the result as shown in Figs. 9 and 10.

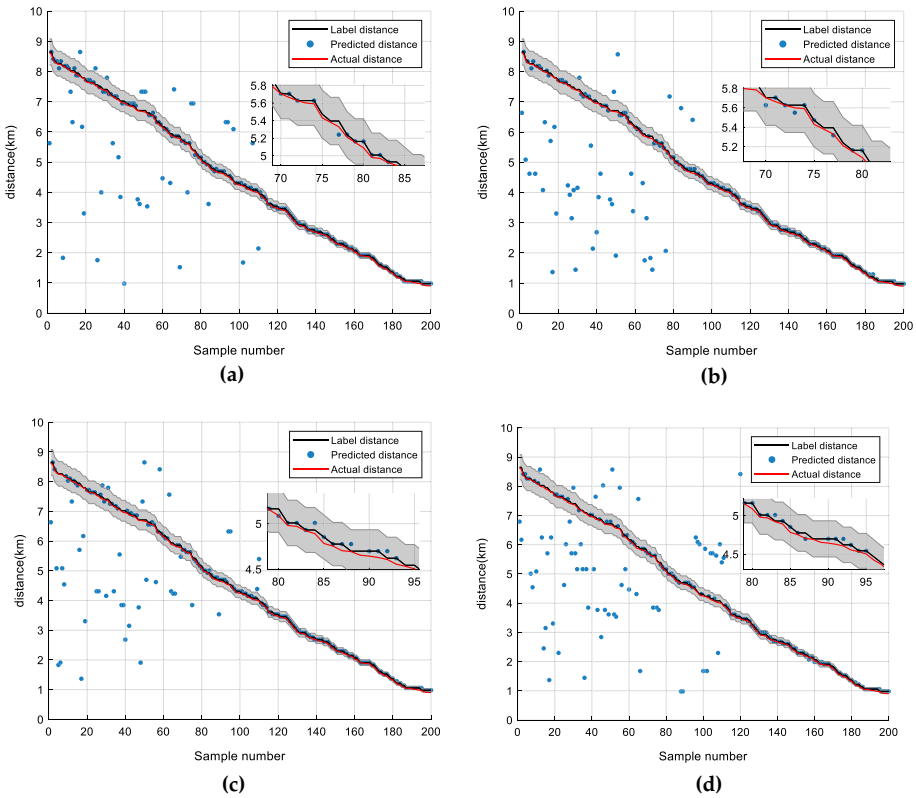

Fig. 9 Prediction plots for the narrowband dataset {163} Hz: **a** RA-CAE-SSL model; **b** CAE-SSL model; **c** RA- SL model; **d** CNN model

From Figs. 9 and 10, it can be seen that the prediction accuracy of each model is high in the range of 1–4 km from the sound source; at greater than 4 km, there are different degrees of outliers in the prediction points, which is because the farther the sound source is, the greater the signal energy attenuation in the propagation process, the lower the signal-to-noise ratio at the receiving end, and the sound signal propagation process is more complex, and the location features are more difficult to extract compared with those at close distances. In addition, it can be found that broadband outliers are less compared with narrowband, and semi-supervised methods have fewer outliers compared with supervision, and the introduction of an attention mechanism can further reduce the outliers.

There also are some limitations in the experiment, such as, we did not use simulation data to verify the effect of the methods, if did, the results may be more credible. In addition, for getting more datasets, the number of snapshot used to get SCM may be not enough, which will reduce the ability to estimate statistical characteristics, and as a result, the model's ability to extract correct features is weakened.

4.3 Source localization performance for the second dataset division method

To verify the generalization ability of the model, the second dataset division method given in Sect. 4.2 is adopted in this subsection, and the data distributions of the training and test sets in the second step are different even in the case of the same label. And we

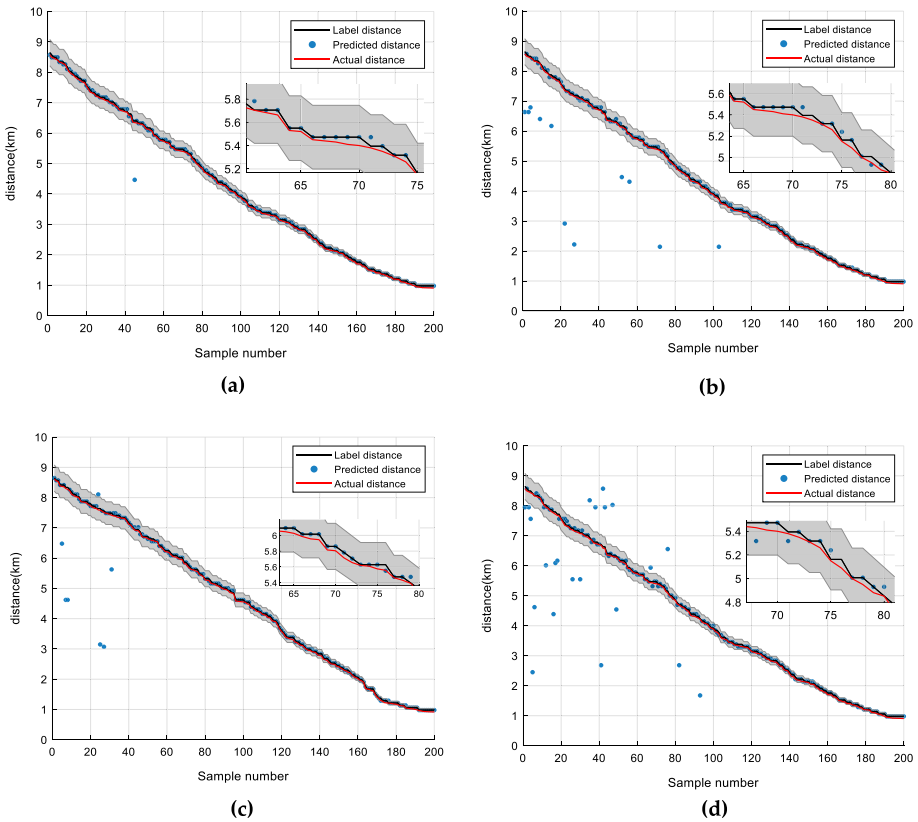

Fig. 10 Prediction plots for the broadband dataset {109, 127, 145, 163, 198, 232, 280} Hz: **a** RA-CAE-SSL model; **b** CAE-SSL model; **c** RA- SL model; (d) CNN model

adjust the number of output channels of the autoencoder convolution layers to find the network with the best generalization ability for the test set in this subsection. The localization performance of the proposed framework with different number of output channels of the autoencoder convolution layers is shown in Table 4.

From the table, it can be seen that the model has the strongest generalization ability to the test set when encoder3 is used. Therefore, the encoder structure of following experiments is referred to encoder3, and the decoder structure is symmetric with the encoder.

The predicted results of the proposed model and the control group are shown in Table 5. The dataset used in the Step 2, which needs label, represents 25% of the entire datasets. And the frequency of dataset is {109,127,145,163,198,232,280} Hz.

The localization performance results in this section show that the proposed framework RA-CAE-SSL has better generalization ability and robustness in the case of "mismatch" between the training and test sets, both compared with the control groups and MFP. As can be seen from Fig. 11a, the prediction points of the RA-CAE-SSL model deviate from the actual distance to different degrees, but they are all close to the actual distance trajectory, and most of the prediction points are within the 5% error limit line, which indicates that the current network model can alleviate the influence of "mismatch"

Table 4 Parameters of the encoder in CAE

	Conv2D_1	Conv2D_2	Conv2D_3	MAE(km)	$P_{CL-5\%}$(%)
Number of output channels(encoder1)	64	128	256	0.0748	59
Number of output channels(encoder2)	64	128	32	0.0666	63
Number of output channels(encoder3)	64	32	16	0.0584	65.5

Table 5 Localization performance of proposed framework and control groups

	MAE (km)	$P_{CL-5\%}$(%)
RA-CAE-SSL	0.0584	65.5
RA-SL	0.0715	58.5
CAE-SSL	0.0731	59
CNN	0.0809	59

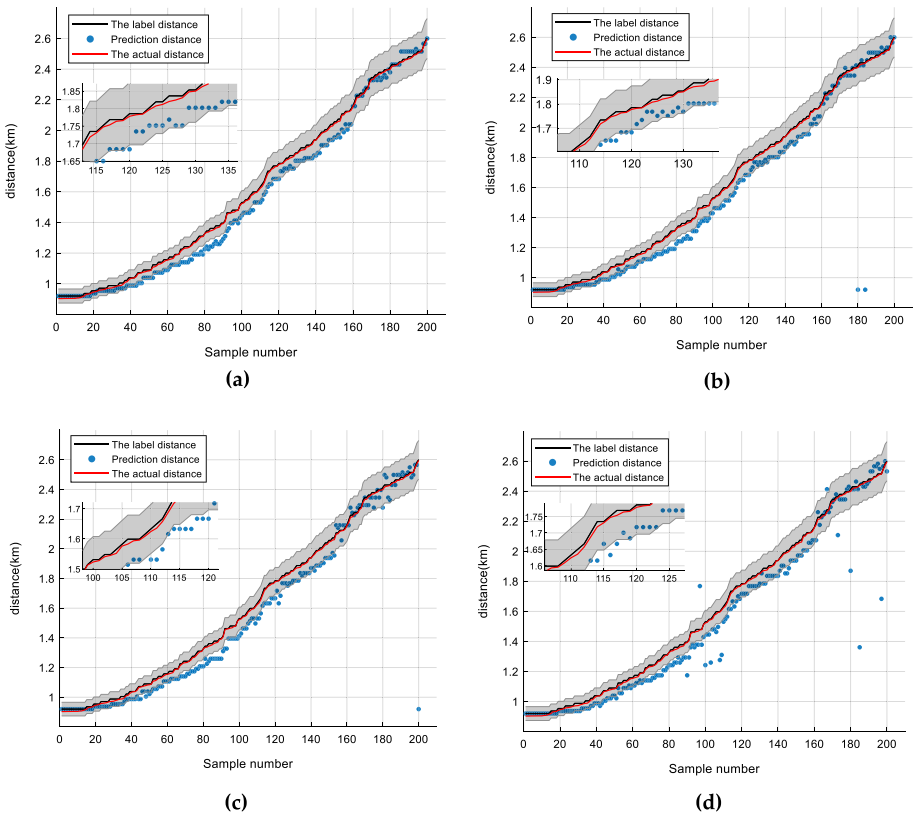

Fig. 11 Prediction results of acoustic source localization: **a** RA-CAE-SSL model; **b** CAE-SSL model; **c** RA- SL model; **d** CNN model

to a certain extent, but cannot fundamentally solve the "The possible reasons for this are: the small amount of data and the limitation that the type of data distribution in our dataset is not rich enough.

In addition, this also shows that the self-encoder can not only learn the complex sound field structure of the shallow sea waveguide, but also summarize the knowledge of the

regularity adapting to different waveguide environments, but the network structure and data richness used in this paper are not enough to further verify this ability of the autoencoder.

5 Conclusion

In this paper, using the idea of semi-supervised learning, the acoustic source localization task is divided into two steps to complete for practical scenarios where labeled data are insufficient and unlabeled data are relatively abundant. A convolutional autoencoder structure incorporating a residual self-attention mechanism module is proposed. The critical feature extraction capability of the autoencoder is effectively improved by SA under the condition where data are affected by noise.

The semi-supervised learning framework for underwater acoustic source localization proposed in this paper is a data-driven approach that, compared to MFP, gets rid of the dependence on precise environmental parameters and theoretically avoids the problem of environmental parameter mismatch. As a consequence, the performance of the proposed framework is clearly better than MFP, but it is not fair to compare them under every scenarios since the prior required by them is different. The data-driven approach mainly depends on the available data: when training data and test data with the same label satisfy the condition of the same distribution, the network model can achieve the best prediction capability; conversely, when the same distribution condition is no longer satisfied, the prediction capability of the model will be significantly affected. It can be seen that the main factor limiting the development of data-driven underwater acoustic source localization methods are the number and quality of data. Based on the data from the SWellEX96 experiment, this paper verifies the localization performance of the proposed method in two cases, and the results show that: 1. The semi-supervised learning framework proposed in this paper is more adaptable to the underwater acoustic field in which limited access to data (minimal access to label data), and the localization performance is more vital than that of supervised learning, especially in the case of the reduced number of label samples, and SA also contributed to it. 2. The network generalization ability of the proposed method is more vital than that of the supervised learning, while has a certain tolerance for differences in data distribution.

It is worth mentioning that the primary purpose of this paper is to demonstrate the advantages of a semi-supervised framework for application in the underwater acoustics. With larger datasets and richer sample data, the framework of this paper can be applied to more complex and powerful networks when better sound source localization performance can be expected to be obtained.

Abbreviations

SSL	Semi-supervised learning
SL	Supervised learning
CAE	Convolutional autoencoder
RA	Residual self-attention mechanism
VLA	Vertical line array
MFP	Matched-field processing
SM	Self-attention mechanism
MSE	Mean square error
CELF	Cross-entropy loss function

MAE Mean absolute error

Acknowledgements
The authors would like to acknowledge the National Natural Science Foundation of China (Grant no. 52071164), the Science and Technology on Underwater Vehicle Technology Laboratory (Grant no.61422152002030)and the Postgraduate Research & Practice Innovation Program of Jiangsu Province (Grant no. KYCX22_3846).

Author contributions
PJ, BW, and LBL contributed to methodology; PJ, BW, and PC contributed to software; PJ and LBL contributed to formal analysis.; PJ, BW, LBL, and FTX contributed to data curation; PJ, BW, and LBL contributed to writing—original draft preparation,; PJ, LBL, PC, and FTX contributed to writing—review and editing; BW contributed to funding acquisition. All authors read and approved the final manuscript.

Funding
This research received no external funding.

Availability of data and materials
Publicly available dataset were analyzed in this study. This data can be found here: http://swellex96.ucsd.edu/index.htm.

Declarations

Ethical approval and consent to participate
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 5 July 2022 Accepted: 19 October 2022
Published online: 08 November 2022

References

1. H.P. Bucker, Use of calculated sound fields and matched field detection to locate sound sources in shallow water. J. Acoust. Soc. Am. **59**(2), 368–373 (1976)
2. R.G. Fizell, S.C. Wales, Source localization in range and depth in an Arctic environment. J. Acoust. Soc. Am. **78**(S1), S57–S58 (1985)
3. E.K. Westwood, Broadband matched-field source localization. J. Acoust. Soc. Am. **91**, 2777–2789 (1992)
4. A.B. Baggeroer, W.A. Kuperman, P.N. Mikhalevsky, An overview of matched field methods in ocean acoustics. IEEE J. Ocean. Eng. **18**(4), 401–424 (1993)
5. Z.H. Michalopoulou, M.B. Porter, Matched-field processing for broadband source localization. IEEE J. Ocean. Eng. **21**, 384–392 (1996)
6. A.B. Baggeroer, Why did applications of MFP fail, or did we not understand how to apply MFP, in *Proceedings of the 1st International Conference and Exhibition on Underwater Acoustics. Corfu, Greece*, 2013, p. 41–9
7. S. Finette, Embedding uncertainty into ocean acoustic propagation models. J. Acoust. Soc. Am. **117**(3), 997–1000 (2005)
8. P. Gerstoft, Inversion of seismoacoustic data using genetic algorithms and a posteriori probability distributions. J. Acoust. Soc. Am. **95**, 770–782 (1994)
9. M. Siderius, P. Gerstoft, P. Nielsen, Broadband geoacoustic inversion from sparse data using genetic algorithms. J. Comput. Acoust. **06**, 117–134 (1998)
10. Z.M. Liu, C.W. Zhang, P.S. Yu, Direction-of-arrival estimation based on deep neural networks with robustness to array imperfections. IEEE Trans. Antennas Propag. **66**(12), 7315–7327 (2018)
11. D. Buscombe, P.E. Grams, Probabilistic substrate classification with multispectral acoustic backscatter: a comparison of discriminative and generative models. Geoscience **8**(11), 395 (2018)
12. N. Allen, P.C. Hines, V.W. Young, Performances of human listeners and an automatic aural classifier in discriminating between sonar target echoes and clutter. J. Acoust. Soc. Am. **130**(3), 1287–1298 (2011)
13. A. Krizhevsky, ImageNet classification with deep convolutional neural networks. Commun. ACM **60**, 84–90 (2012)
14. T.L. Hemminger, Y.H. Pao, Detection and classification of underwater acoustic transients using neural networks. IEEE Trans. Neural Netw. Learn. Syst. **5**(5), 712–718 (1994)
15. J. Choi, Y. Choo, K. Lee, Acoustic classification of surface and underwater vessels in the ocean using supervised machine learning. Sensors **19**(16), 3492 (2019)
16. J. Chi, X. Li, H. Wang, Sound Source Ranging Using a Feed-forward Neural Network with Fitting-based Early Stopping. J. Acoust. Soc. Am. **146**(3), EL258–EL264 (2019)
17. H. Niu, E. Reeves, P. Gerstoft, Source localization in an ocean waveguide using supervised machine learning. J. Acoust. Soc. Am. **142**(3), 1176–1188 (2017)
18. X. Wang, A. Liu, Y. Zhang, Underwater acoustic target recognition: a combination of multi-dimensional fusion features and modified deep neural network. Remote Sens. **11**(16), 1888 (2019)
19. K.M. Martin, W.T. Wood, J.J. Becker, A global prediction of seafloor sediment porosity using machine learning. Geophys. Res. Lett. **42**(24), 10640–10646 (2015)
20. J.C. Park, R.M. Kennedy, Remote sensing of ocean sound speed profiles by a perceptron neural network. IEEE J. Ocean. Eng. **21**(2), 216–224 (1996)

21. M. Bianco, P. Gerstoft, Dictionary learning of sound speed profiles. J. Acoust. Soc. Am. **141**(3), 1749–1758 (2017)
22. Y. Mahmutoglu, K. Turk, E. Tugcu, Particle swarm optimization algorithm based decision feedback equalizer for underwater acoustic communication, in *2016 39th International Conference on Telecommunications and Signal Processing (TSP)*, 2016, p. 153–156
23. Yu. Jing Zhang, X.F. Cao, Deep neural network-based underwater OFDM receiver. IET Commun. **13**, 1998–2002 (2019)
24. Y. Chen, Yu. Weijian, X. Sun et al., Environment-aware communication channel quality prediction for underwater acoustic transmissions: a machine learning method. Appl. Acoust. **181**, 108128 (2021)
25. Su. Yuhan, M. Liwang, Z. Gao et al., Optimal cooperative relaying and power control for IoUT networks with reinforcement learning. IEEE Internet Things J. **8**, 791–801 (2021)
26. Z. Jin, Q. Zhao, Su. Yishan, RCAR: a reinforcement-learning-based routing protocol for congestion-avoided underwater acoustic sensor networks. IEEE Sens. J. **19**, 10881–10891 (2019)
27. Y. Chen, K. Zheng, X. Fang et al., QMCR: A Q-learning-based multi-hop cooperative routing protocol for underwater acoustic sensor networks. China Commun. **18**, 224–236 (2021)
28. S. Wei, J. Lin, K. Chen, Reinforcement learning-based adaptive modulation and coding for efficient underwater communications. IEEE Access **7**, 67539–67550 (2019)
29. H. Yang, K. Lee, Y. Choo, Underwater acoustic research trends with machine learning: general background. IEEE J. Ocean. Eng. **34**(2), 147–154 (2020)
30. H. Niu, Z. Gong, E. Ozanich, Deep-learning source localization using multi-frequency magnitude-only data. J. Acoust. Soc. Am. **146**, 211–222 (2019)
31. W. Wenbo, N. Haiyan, S. Lin, H. Tao, Deep transfer learning for source ranging: Deep-sea experiment results. J. Acoust. Soc. Am. **146**(4), EL317–EL322 (2019)
32. J. Wang, R. Fan, Underwater target tracking method based on convolutional neural network, in *2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA)*, 2021, p. 636–640
33. J. Murray, D. Ensberg, The Swellex-96 Experiment, 1996. Available online: http://www.swellex96.ucsd.edu/index.htm
34. M.S. Seyfioğlu, A.M. Özbayoğlu, S.Z. Gürbüz, Deep convolutional autoencoder for radar-based classification of similar aided and unaided human activities. IEEE Trans. Aerosp. Electron. Syst. **54**, 1709–1723 (2018)
35. C. Min, S. Xiaobo, Z. Yin, Deep feature learning for medical image analysis with convolutional autoencoder neural network. IEEE Trans. Big Data **07**, 750–758 (2021)
36. P.Y. Wang, C.T. Chen, S.H. Huang, Deep learning model for house price prediction using heterogeneous data analysis along with joint self-attention mechanism. IEEE Access **09**, 55244–55259 (2021)
37. Q. Wang, Z. Teng, J. Xing, Learning attentions: residual attention siamese network for high performance online visual tracking, in *CVF Conference on Computer Vision and Pattern Recognition, IEEE Computer Society*, 2018, p. 4854–4863
38. S. Noh, D.J. Ji, D.H. Cho, A self-attention-based I/Q imbalance estimator for beyond 5G communication systems. IEEE Commun. Lett. **25**, 3262–3266 (2021)
39. Y. Qian, J. Qi, X. Kuai, Specific emitter identification based on multi-level sparse representation in automatic identification system. IEEE Trans. Inf. Foren Sec. **16**, 2872–2884 (2021)

Publisher's Note