remote sensing

MDPI

Article

MTCL-Net: A Multi-Task Contrastive Learning Network for Underwater Acoustic Source Ranging

Jixiang Zhao [1,2,3], Zhiliang Qin [1,2,3,4,5,*], Benjun Ma [1,2,3,4], Wenjian Lan [1,2,3,4], Bingqi Liu [4] and Shuyi Pang [1,2,3]

[1] National Key Laboratory of Underwater Acoustic Technology, Harbin Engineering University, Harbin 150001, China; zhaojixiang@hrbeu.edu.cn (J.Z.); mabenjun@hrbeu.edu.cn (B.M.); lanwj@hrbeu.edu.cn (W.L.); psy@hrbeu.edu.cn (S.P.)
[2] Key Laboratory of Marine Information Acquisition and Security, Harbin Engineering University, Ministry of Industry and Information Technology, Harbin 150001, China
[3] College of Underwater Acoustic Engineering, Harbin Engineering University, Harbin 150001, China
[4] Qingdao Innovation and Development Center, Harbin Engineering University, Qingdao 266400, China; liubingqi@hrbeu.edu.cn
[5] Sanya Nanhai Innovation and Development Base, Harbin Engineering University, Sanya 572000, China
* Correspondence: qinzhiliang@hrbeu.edu.cn

Highlights

What are the main findings?

- A novel multi-task contrastive learning network (MTCL-Net) is proposed for underwater acoustic source ranging, with a Siamese contrastive learning auxiliary task for source position similarity discrimination.
- The proposed MTCL-Net framework realizes a mean absolute error of 0.17 km and 90.36% probability of credible localization (PCL-10%) on the SWellEx-96 sea trial dataset and reaches an optimal ranging performance with only around 60% of measured samples for fine-tuning.

What are the implications of the main findings?

- This work overcomes the strong dependence of classical underwater source localization methods on precise marine environmental parameters, providing a novel solution for passive acoustic ranging in complex and uncertain ocean environments.
- The contrastive learning-enhanced multi-task learning paradigm established in this study offers a generalizable research path for few-shot learning and environmental mismatch mitigation in underwater acoustic signal processing.

Abstract

Deep learning-based data-driven methods have gained significant attention in underwater acoustic source localization. However, their performance is often constrained by environmental disturbances and the scarcity of real-world underwater acoustic data. To address these issues, this paper presents a novel network termed MTCL-Net, a multi-task learning network that incorporates contrastive learning as an auxiliary task for underwater acoustic source ranging. A standard dataset and a perturbed dataset to simulate real underwater interferences are constructed based on known environmental parameters in this method. A Siamese dual-branch architecture is employed, where a contrastive learning task enables the automatic extraction of position-related features. The network jointly optimizes three tasks: source localization in perturbed environments, localization on the standard dataset, and position similarity discrimination, which improves the robustness and generalization ability. The experimental results on simulated and sea trial data demonstrate that MTCL-Net outperforms traditional matched field processing (MFP), single-task learning

Check for updates

Academic Editors: Gang Zheng, Lei Ren and Indiana A. Olbert

Received: 8 March 2026
Revised: 18 April 2026
Accepted: 22 April 2026
Published: 27 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

(STL), and multi-task learning based on depth–range (MTL-DR) methods in terms of mean absolute error (MAE) and probability of credible localization (PCL-10%). Specifically, on SWellEx-96 sea trial data, MTCL-Net achieves an MAE of 0.17 km and a PCL-10% of 90.36%. Moreover, the proposed method only needs a few samples for fine-tuning and shows strong practicality in uncertain marine environments.

Keywords: underwater acoustic source ranging; multi-task learning; contrastive learning; source position similarity discrimination; SWellEx-96

1. Introduction

Passive source localization using hydrophone arrays has long been a core research topic in underwater acoustics, motivated by its essential applications in underwater security and marine surveillance. As a typical model-based method, matched field processing (MFP) integrates underwater acoustic propagation models with array signal processing techniques [1–3]. Despite decades of extensive study and numerous valuable contributions, the performance of MFP remains inherently limited by its sensitivity to mismatches between assumed and actual environmental parameters [4–7].

In recent years, data-driven methodologies, particularly deep learning (DL), have garnered significant attention for underwater source ranging and depth estimation [8–11]. A growing body of literature has demonstrated that DL-based methods can achieve a superior localization performance compared to classical MFP [12–16]. Unlike model-based techniques, DL models excel at extracting complex, nonlinear features from the acoustic field that correlate with source position, effectively learning statistical patterns directly from the data. This advantage makes deep learning a promising paradigm for source localization in uncertain underwater environments [12,17,18]. However, the scarcity of labeled underwater acoustic data remains a critical bottleneck restricting the further development and practical deployment of DL in this field. To alleviate this challenge, training models on simulated data has emerged as a core, widely adopted strategy [12,14,17,19–21]. By constructing simulated acoustic field datasets covering diverse combinations of environmental parameters, we can substantially expand the scale of training data [12], enabling the development of deeper and more sophisticated models with enhanced feature extraction capabilities. Furthermore, exposing models to a wide spectrum of environmental perturbations during training can effectively boost their robustness. Nevertheless, this approach incurs a significant computational overhead due to the massive number of parameter combinations required for dataset construction. More critically, discrepancies between simulation models and real ocean environments—stemming from inherent modeling inaccuracies and environmental parameter mismatches—often lead to a poor generalization performance when models trained exclusively on simulated data are applied to real field measurements [8].

To address these data-related challenges, a variety of techniques have been investigated. Semi-supervised learning with automatic sample labeling has been introduced for underwater source ranging [22,23]. Approaches such as autoencoders and contrastive learning generate pseudo-labels for large unlabeled datasets [24,25], thus expanding the training set and reducing the dependence on the manual annotation of measured data. In addition, unsupervised domain adaptation (UDA) has also been utilized to bridge the domain gap between simulated and measured data by aligning their feature distributions, yielding a promising performance even when target domain range labels are unavailable [26,27]. However, these methods still demand a considerable volume of measured data (even if

unlabeled) to capture the target distribution or enrich the training corpus. Transfer learning, which performs pre-training on large-scale simulated data followed by fine-tuning with limited measured samples [19], has emerged as an effective and widely adopted solution. A key challenge within this paradigm lies in achieving efficient model adaptation using minimal real data to ensure practicality, while preserving robustness against the complex and uncertain perturbations inherent in real oceanic environments. Beyond data-centric strategies, the architectural design of deep learning models plays a pivotal role. Various network structures, including deep neural networks (DNNs) [28,29], convolutional neural networks (CNNs) [17,20,30,31], and residual networks (ResNet) [12,32,33], have been widely applied to underwater source localization. Among these, multi-task learning (MTL) has attracted growing interest. Since source range and depth are typically physically coupled, MTL provides a natural framework for their joint estimation. Studies have shown that MTL-based approaches achieve superior localization performances over conventional single-task learning by exploiting shared feature representations [15,17,21,34]. Furthermore, several works have adopted MTL to jointly estimate environmental parameters and source positions, introducing additional physical constraints that improve model stability and generalization [35–37]. The demonstrated merits of such MTL frameworks offer meaningful guidance for integrating richer constraints into localization models, which motivates the method proposed in this paper.

Inspired by the related studies, this paper proposes a novel multi-task learning framework (MTCL-Net) that embeds source spatial location constraints for underwater acoustic source range estimation. Specifically, the main contributions of this work are summarized as follows:

(1) To address real ocean environmental uncertainty without a sharp increase in training data scale, we build an environmental perturbed dataset by randomizing key environmental parameters, which is paired with a standard dataset under fixed nominal parameters to characterize real ocean environmental variability.
(2) Based on a Siamese dual-branch architecture, we design an acoustic source position similarity discrimination task via contrastive learning between the standard and perturbed datasets, which enabled the model to automatically learn position-related discriminative features and improve robustness to environmental mismatches.
(3) A multi-task learning mechanism is designed to jointly optimize the primary localization tasks on both datasets and the auxiliary contrastive task. This joint training strengthens inter-task constraints and effectively enhances the model's overall localization performance and environmental robustness.

2. Materials and Methods

2.1. Overview of the Proposed MTCL-Net Framework

The workflow of the proposed MTCL-Net framework (as shown in Figure 1) is systematically elaborated in this section, organized into three sequential core stages: dataset construction, network architecture design, and model training with field data adaptation. Details are described as follows:

Firstly, paired standard and perturbed acoustic field datasets for model training are constructed based on the environmental parameters of the classic SWellEx-96 sea trial experiment [38]. Specifically, the KRAKEN normal mode acoustic propagation model [39] is adopted to generate sound pressure data at multiple preset source locations under fixed deterministic environmental parameters, which forms the standard dataset. On this basis, random perturbations within pre-defined reasonable ranges are introduced to key marine environmental parameters, including in-water sound speed, seabed sound speed, seabed attenuation and density, water depth, and seabed sediment layer thickness. These

randomized perturbed combinations are used to generate sound pressure data at the same source locations to simulate the impacts of real ocean environmental uncertainty on the acoustic field, with the resulting data constituting the perturbed dataset.

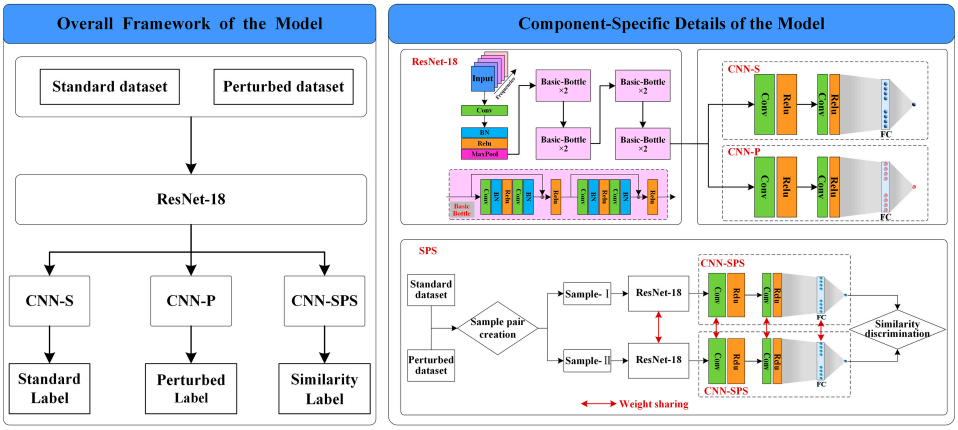


Figure 1. MTCL-Net framework.

Secondly, the multi-task learning architecture of MTCL-Net with a Siamese dual-branch structure is designed. ResNet-18 [40] is adopted as the shared backbone for general acoustic field feature extraction, on top of which three task-specific branches are built: the CNN branch trained on the standard dataset (CNN-S), the CNN branch trained on the perturbed dataset (CNN-P), the source position similarity (SPS) discrimination between perturbed and standard environments, and their corresponding CNN branches (CNN-SPS). Among these, CNN-S and CNN-P serve as the primary source localization tasks, while CNN-SPS acts as the auxiliary contrastive learning task. All three branches share the ResNet-18 feature encoder to ensure consistent feature representation across tasks. This multi-task joint learning paradigm strengthens inter-task consistency constraints, and significantly enhances the backbone's capability to extract discriminative, position-sensitive acoustic field features.

Thirdly, a two-stage training and fine-tuning strategy is developed to improve the model's practicality in real marine environments. After pre-training of the full MTCL-Net framework is completed on the simulated paired datasets, the parameters of the shared ResNet-18 backbone are frozen to preserve the generalizable acoustic feature extraction capability learned during pre-training. Subsequently, lightweight fine-tuning is performed exclusively on the task-specific branches using a small amount of real measured sea trial data to achieve efficient domain adaptation and a reliable localization performance on real-world field data.

2.2. Vertical Array Signal Preprocessing

Raw array signals are typically represented in the time domain. In underwater acoustic signal processing and analysis, converting time domain signals into frequency domain signals is a common practice to reduce computational complexity and effectively capture varying frequency responses. In this study, the time domain signals are transformed into frequency responses via Fourier transform, and subsequently, a sound pressure covariance matrix is constructed to serve as the model's input. The specific steps are as follows.

Sound pressure data can be represented as $p = [p_1, p_2, \cdots, p_f, \cdots, p_F]$, where p_F indicates the sound pressure data extracted at frequency F after Fourier transforming the time domain signal. In underwater acoustic field theory models, the sound pressure p_f at frequency f is expressed as

$$p_f = S(f)g(f, r) + \varepsilon \quad (1)$$

where $S(f)$ represents the source term, $g(f,r)$ is the acoustic field excited by a point source, influenced by the source frequency, ocean environment, and source position, expressed as a Green's function, r indicates the source distance, and ε represents noise.

Let the number of hydrophones in the vertical linear array be L. The complex sound pressure obtained from the hydrophone array can be expressed as

$$p(f) = [p_1(f), p_2(f), \cdots, p_L(f)]^{\mathrm{T}}. \tag{2}$$

To reduce differences between signals, the complex sound pressure is normalized using the L_2 norm, and the process can be represented as

$$\widehat{p}(f) = \frac{p(f)}{\|p(f)\|_2} = \frac{[p_1(f), p_2(f), \cdots, p_L(f)]^{\mathrm{T}}}{\sqrt{\sum_{i=1}^{L} |p(f)|^2}} \tag{3}$$

The sound pressure covariance matrix can be calculated using the following formula:

$$C(f) = \widehat{p}(f) \cdot \widehat{p}(f)^{\mathrm{H}} \tag{4}$$

where $C(f)$ is the covariance matrix (SCM) of $L \times L$. The real and imaginary parts of the matrix are separated; for single-frequency signals, the model features can be obtained as $L \times L \times 2$. For broadband signals, SCMs at different frequencies are stacked along the third dimension, resulting in feature input dimensions of $L \times L \times 2F$, where F is the number of frequencies.

2.3. Siamese Network Based on Acoustic Source Spatial Position Similarity

The sound field in an ocean waveguide is coupled with parameters of the acoustic source (frequency and position) and the environment (sound speed profile, seabed and sea surface properties) through a complex nonlinear process. The model proposed in this study aims to estimate the acoustic source distance by leveraging the powerful nonlinear feature extraction ability of deep learning methods, while avoiding the adverse effects introduced by environmental perturbed factors as much as possible. Based on the constructed "standard dataset" and "perturbed dataset", the model we proposed constructs a Siamese network module based on the similarity measure of the acoustic source spatial position (which also is the CNN-SPS in Figure 1). We selected sound field samples with the same spatial position in the "standard dataset" and the "perturbed dataset" and carried out iterative training using contrastive learning to achieve the focus of the model on the features related to the acoustic source position. The design idea based on the similarity measure of the acoustic source spatial position is shown in Figure 2.

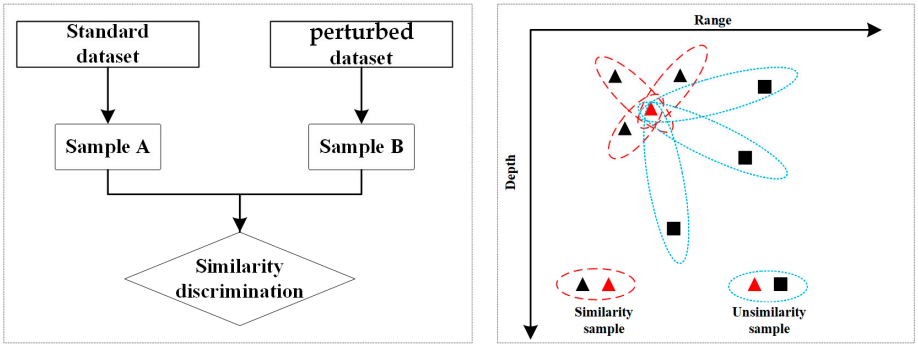


Figure 2. The design idea based on the similarity measure of the acoustic source spatial position.

In this model, the CNN-SPS module is the core that connects the "standard dataset" and the "perturbed dataset" to achieve the acoustic source spatial position constraint. We adopt a two-branch Siamese network structure to conduct iterative learning on samples from different datasets. The Siamese neural network is a deep learning architecture for contrastive learning. Through a two-channel symmetric structure (sub-networks with shared parameters), this network can map input samples to the same high-dimensional feature space and discriminate the similarity of samples based on the distance metric of the embedding vectors, that is, to compare whether the distance labels of the input samples are the same. This network consists of 2 sub-networks with shared weights. It can process 2 inputs simultaneously (i.e., the 2 samples in the input sample pair), extract features of the input samples through a CNN, map the high-dimensional features of the 2 input samples to the same dimensional space using a fully connected structure, and judge their similarity by comparing the sample distances in the feature space. Since the Siamese neural network usually processes paired input data and optimizes the network weights by comparing the similarity between two samples, in this study, sample pairs need to be constructed first according to the consistency of the underwater acoustic source distance (acoustic source distances that are the same are called "positive sample pairs", and those that are different are called "negative sample pairs"). To enhance the model's ability to extract acoustic source spatial position features, in this study, samples from different datasets within a certain spatial neighborhood range are regarded as "positive sample pairs", and samples from different datasets outside this spatial neighborhood range are regarded as "negative sample pairs". As is shown in Figure 3, the position in the shadow and the acoustic source position (red pentagram) form "positive sample pairs", while the other positions and the acoustic source position form "negative sample pairs".

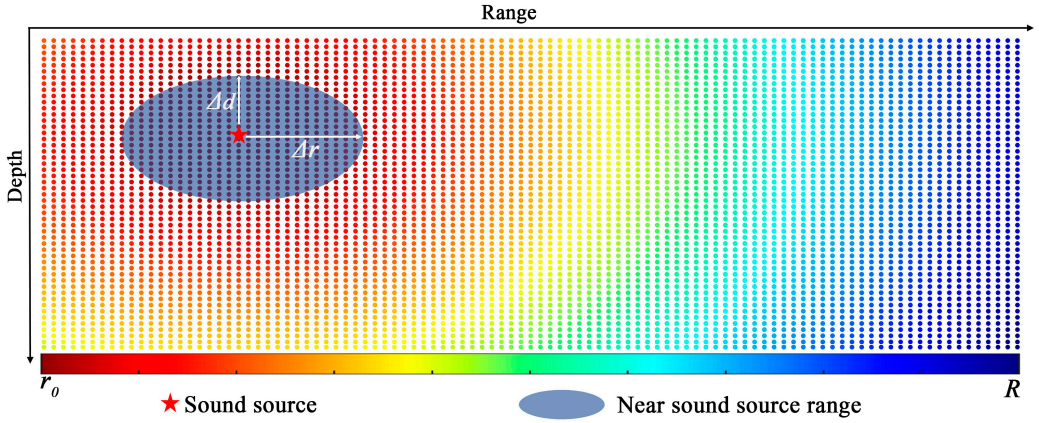


Figure 3. The construction of positive and negative sample pairs.

The construction of positive and negative sample pairs follows the steps below:

(1) Randomly select a sample s_{si} in the "standard dataset" whose acoustic source position is (d_{si}, r_{si});

(2) Select a sample s_{pi} in the "perturbed dataset" whose acoustic source position is (d_{pi}, r_{pi});

(3) As shown in the figure, if it is determined that the position of the s_{pi} acoustic source is within the neighborhood range (shaded area) of the s_{si} acoustic source position, then s_{si} and s_{pi} form a positive sample pair; otherwise, the two form a negative sample pair, that is

$$\begin{cases} d_{pi} < |d_{si} \pm \Delta d| \&\& d_{ri} < |r_{si} \pm \Delta r|, & y_{spsi} = 1 \\ otherwise, & y_{spsi} = 0 \end{cases} \tag{5}$$

where y_{spsi} is the tag value of the constructed sample pair, y_{spsi} = 1 means a positive sample pair, y_{spsi} = 0 means a negative sample pair, and Δd is the search interval in the depth direction, while Δr is in the range direction. In order to maintain the balance of positive and negative sample pairs, the sampling ratio of positive and negative samples can be set to 50%.

In this paper, the contrast loss function is used to measure the feature similarity. The contrast loss function [41] is as follows:

$$Loss_{sps} = \frac{1}{2N}\sum_{i=1}^{N}\left(y_{spsi} \cdot D_i^2 + (1 - y_{spsi}) \cdot \max(margin - D_i, 0)^2\right) \tag{6}$$

where y_i is the label, 1 is the positive sample pair (the distance labels of two samples in the sample combination are consistent), and 0 is the negative sample pair (the distance labels of two samples in the sample combination are inconsistent). D_i is the Euclidean distance of the sample pair, *margin* is a superparameter which represents the minimum distance between negative sample pairs. Therefore, the loss function can better describe the matching degree between sample pairs as follows:

When two samples are similar or matched, $y_i = 1$, the loss function is $L = \frac{1}{2N}\sum_{i=1}^{N} y_i \cdot D_i^2$. If the Euclidean distance of the extracted feature results is large, the loss output by the network at the end is large, indicating that the model still needs further training.

When the two samples are not similar or do not match, $y_i = 0$, the loss function is $L = \frac{1}{2N}\sum_{i=1}^{N}(1 - y_i) \cdot \max(margin - D_i, 0)^2$. If the Euclidean distance of the extracted feature results is small, the final output loss of the network will increase, indicating that the model still needs further training.

The model trained by the above iterative optimization method can automatically determine whether the input sample pairs have the same acoustic source distance label. Since the convolutional structure in this network has the same weights, it can be regarded as performing the same feature mapping operation on the two input samples: when processing positive sample pairs (same range label), the features extracted by the network are highly similar; for negative sample pairs (different range labels), their feature representations show significant differences. During the training process, the network dynamically adjusts the feature extraction weights to make the model focus on learning the key discriminative features of the acoustic source distance estimation task. This mechanism makes the feature representations extracted by the network have a strong correlation with the acoustic source distance, and at the same time helps to suppress interference factors, enabling the model to have the ability to decouple and predict the acoustic source position information in the sound field. In addition, the shared weight and structure design of this Siamese network can reduce the number of model parameters, which helps to improve the robustness of the model.

The feature extraction network designed in this paper is trained entirely based on the simulation data generated by the sound field model. Although deep networks have powerful feature extraction and fitting capabilities, considering the limited amount of measured underwater acoustic data, to avoid the domain adaptation ability of the model decreasing due to overfitting the simulation data in the actual scenario, the number of network layers for feature extraction should not be too many. Therefore, this study adopted a streamlined design with only two convolutional layers. This lightweight network design

can reduce the model complexity and improve the training speed while retaining the necessary feature abstraction ability.

2.4. Multi-Task Learning Distance Estimation Based on Spatial Position Constraint of Acoustic Source

This research uses the array acoustic pressure covariance matrix as the model input. Considering that its two-dimensional matrix characteristics are naturally adaptable to the local receptive field mechanism of a CNN, this paper constructs a feature extraction network with a CNN as the core. The overall architecture of the model is shown in the figure. Considering the training cost, this paper directly uses the basic architecture of the ResNet-18 network as the common feature extraction structure for different datasets. The residual network is a classic deep learning network model that solves the problem of vanishing or exploding gradients due to the increase in the number of model layers. By designing a unique skip connection structure to construct residual blocks, the residual network can stack a large number of residual blocks, thereby realizing the extraction and utilization of deep features of the model and improving the model performance. In this study, the ResNet-18 network contains a total of 1 independent 7×7 convolutional layer and 8 residual blocks, where each residual block has two 3×3 convolutional layers, and the residual blocks form an "identity mapping" through skip connections.

MTL improves the generalization ability by using the specific domain information contained in the training signals of related tasks. In deep learning, MTL processes the hidden layer by sharing parameters. Due to the adoption of the MTL method, the output stream of ResNet-18 consists of three branches, corresponding to the distance estimation of samples in the "standard dataset" (CNN-S), the estimation of samples in the "perturbed dataset" (CNN-P), and the output of the similarity of the acoustic source spatial position (CNN-SPS). Constraints can be formed between different tasks to further improve the expression of the model for acoustic source distance-related features and enhance the model performance. Each task contains 2 convolutional layers and 2 fully connected layers. The sizes of the 2 convolutional layers are both 3×3; the first fully connected layer has 256 neurons, and the second fully connected layer has 1 neuron for direct output.

Among them, MSE is used as the loss function for both the distance estimation of samples in the "standard dataset" and the estimation of samples in the "perturbed dataset", that is

$$Loss_s = \frac{1}{N}\sum_{i=1}^{N}(y_{si} - \hat{y}_{si})^2 \tag{7}$$

$$Loss_p = \frac{1}{N}\sum_{i=1}^{N}(y_{pi} - \hat{y}_{pi})^2 \tag{8}$$

where y_{si} is the distance tag corresponding to N samples in the standard data set, $\hat{y}_{si}$ is the distance value predicted in CNN-S, y_{pi} is the distance tag corresponding to N samples in the disturbed data set, and $\hat{y}_{pi}$ is the distance value predicted in CNN-P.

Its loss function is

$$Loss = Loss_s + Loss_p + \alpha Loss_{sps} \tag{9}$$

where α is the scale factor, which can keep the losses of $Loss_s$, $Loss_p$ and $Loss_{sps}$ to similar amplitude values to accelerate the model iteration training speed. Through experimental tests, α can be set to 0.1 in this study.

2.5. Performance Metrics

This paper uses three objective indicators, namely the mean absolute error (MAE) [42] and the probability of credible localization (PCL) [15], to measure the distance estimation performance of different methods. The calculation method of the MAE is as follows:

$$MAE = \frac{1}{M}\sum_{i=1}^{M}|r_i - \hat{r}_i| \tag{10}$$

The calculation method of the PCL for distance estimation is as follows:

$$PCL = \frac{1}{M}\sum_{i=1}^{M}\eta_i \times 100\%, \eta_i = \begin{cases} 1 & if \left|\frac{\hat{r}_i - r_i}{r_i}\right| \leq 0.1 \\ 0 & otherwise \end{cases} \tag{11}$$

where M represents the number of test samples, $\hat{r}_i$ is the predicted value of the sample distance, and r_i is the true value of the sample distance. The probability of credible localization (PCL) for distance estimation is defined as the proportion of samples whose distance estimation error is within 10% and is denoted as PCL-10%.

3. Experimental Results and Analysis

3.1. Simulation and Data Set

This study generates a "standard dataset" based on the environmental parameters provided by the SWellEx-96 experiment. The simulated environment is illustrated in Figure 4, with the water layer depth set at 216.5 m and a density of 1 g/cm^3. Beneath the water layer lies a 23.5 m-thick sedimentary layer with a density of 1.76 g/cm^3 and an attenuation of 0.2 dB/km Hz. The top and bottom sound speeds are 1572.3 m/s and 1593.0 m/s, respectively. Below the sedimentary layer is an 800 m-thick mudstone layer with a density of 2.061 g/cm^3 and an attenuation of 0.06 dB/km Hz. The top and bottom sound speeds of this layer are 1881 m/s and 3245 m/s, respectively. The seabed consists of a semi-infinite space with a density of 2.661 g/cm^3, an attenuation of 0.02 dB/km Hz, and a sound speed of 5200 m/s.

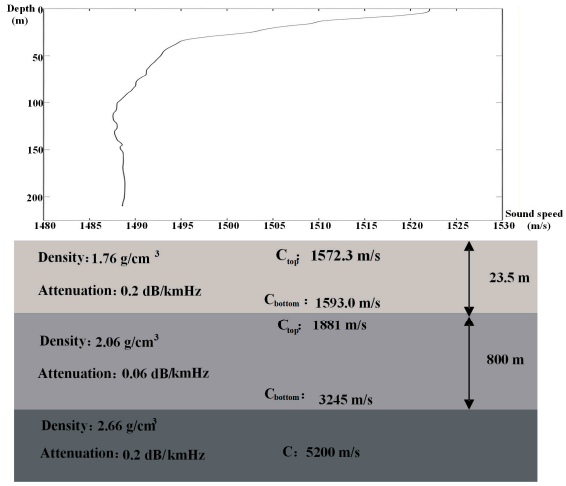


Figure 4. The environmental parameters provided by the SWellEx-96 experiment.

The above marine environment parameters are used as the reference data, and random disturbances are added according to the range shown in Table 1. The "perturbed dataset" in the simulation environment is generated by a random combination of different parameters.

Table 1. Acoustic field simulation parameters in perturbed dataset.

Parameter	Benchmark Parameters	Disturbance Parameters
Seawater sound velocity	data	±5%
Sedimentary layer sound Velocity-1	1572.3–1953.0 m/s	±5%
Sedimentary layer sound Velocity-2	1881–3245 m/s	±5%
Density of sedimentary layer-1	1.76 g/cm^3	±5%
Density of sedimentary layer-2	2.06 g/cm^3	±5%
Sedimentary layer attenuation-1	0.2 dB/km Hz	±0.05 dB/km Hz
Sedimentary layer attenuation-2	0.06 dB/km Hz	--
Sedimentary layer thickness-1	23.5 m	±2.35 m
Sedimentary layer thickness-2	800 m	±20 m

According to the parameter settings given by the SWellEx-96 experiment, the simulation frequency in this paper is {49, 64, 79, 94, 112, 130, 148, 166, 201, 235, 283, 338} Hz in the marine environment. The vertical hydrophone receiving array has a total of 21 elements, and the depth of the elements is referred to. A total of 40 groups of acoustic source depth are taken from 5 m to 200 m at an interval of 5 m. A total of 901 groups are taken from 1 km to 10 km at an interval of 10 m from the array to the acoustic source. There are $40 \times 901 = 36{,}040$ samples in the standard data set and disturbance data set respectively. Since the covariance of sound pressure is a complex matrix, the real part and imaginary part are separated in the data set construction stage, and each sample dimension formed by the construction is $21 \times 21 \times 24$. According to the generation method of the disturbance data set, the test set with completely independent environmental parameters is generated. The acoustic source depth is taken from 5 m to 200 m at an interval of 5 m, a total of 40 groups, and the array distance from the acoustic source is taken from 1 km to 10 km at an interval of 100 m, giving a total of 91 groups. The independent test data set contains 3640 samples to verify the effectiveness of the analysis model.

All experiments were implemented in the PyTorch 2.11.0 framework. The model was trained with the Adam optimizer, using a batch size of 200, a momentum coefficient of 0.9, an initial learning rate of 0.001, and a learning rate decay factor of 0.1. To guarantee the fairness of the comparison experiments, all benchmark methods employed in this work were trained using the same set of parameters.

3.2. Simulation Results and Analysis

3.2.1. Simulation Results

To verify the effectiveness of the MTCL-Net method, the model is first trained on simulated datasets consisting of both the standard dataset and the perturbed dataset, and then evaluated on an independent test dataset. Comparative analyses are conducted with the matched field processing (MFP) method, the depth–range-based multi-task learning method (MTL-DR), and the conventional single-task learning (STL) method. Notably, both the MTL-DR and STL methods are trained on the combined dataset of the standard and perturbed environments. For intuitive visualization, range estimates and corresponding ground truths at water depths of 25 m, 75 m, 125 m, and 175 m are selected from the independent test dataset, with the results displayed in the figures. Compared with MFP, MTL-DR, and STL, MTCL-Net achieves substantially better consistency with the ground truth, and most predictions fall within the 10% relative distance error range. The distance

estimation results of the MFP method are presented in Figure 5b. Owing to the existence of random environmental perturbations in the test dataset, MFP fails to provide an accurate ranging performance under such conditions. As illustrated in Figure 5c,d, the prediction results of MTL-DR and STL respectively show an overall trend consistent with the true distances. Nevertheless, considerable deviations from the ground truth are observed at certain ranges, demonstrating that these methods still suffer from performance degradation under environmental perturbations.

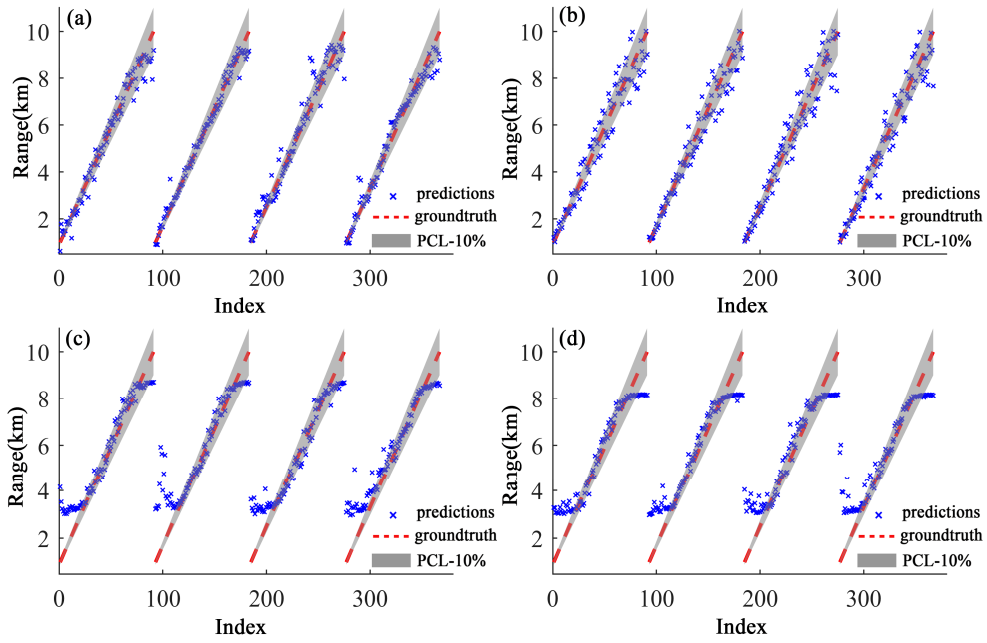


Figure 5. The result of range estimates in test dataset. (**a**) The range estimates of MTCL-Net method; (**b**) the range estimates of MFP method; (**c**) the range estimates of MTL-DR method; (**d**) the range estimates of STL method.

Table 2 summarizes the mean absolute error (MAE) and probability of credible localization (PCL-10%) of all compared methods on the entire independent test dataset. Unlike conventional test setups, the test dataset used herein is constructed from acoustic fields in marine environments distinct from those of the training data, presenting inherent cross-domain characteristics that enable a more rigorous evaluation of the model's robustness and generalization performance. The conventional matched field processing (MFP) method, as a fully physics-driven localization approach, achieves an MAE of 0.5 km on the perturbed test dataset generated around the nominal standard marine environment. However, its PCL-10% is only 56.90%, which demonstrates that its performance is severely impaired by marine environmental uncertainty, with non-negligible large deviations in partial range estimation results. The MTL-DR and STL methods are both data-driven approaches. These two methods struggle to achieve perfect alignment between predicted ranges and ground truth values, resulting in relatively large MAEs. Nevertheless, both yield improved PCL-10% performances compared with the MFP method. Notably, MTL-DR achieves significantly more accurate range estimation than STL, which clearly verifies the inherent advantages of the multi-task learning paradigm. The proposed MTCL-Net achieves the optimal overall performance on both MAE and PCL-10% metrics, outperforming all baseline methods. This result fully validates the effectiveness of the multi-task learning framework with acoustic source spatial position constraints established in this work.

Table 2. Performance comparison of different methods on the test dataset.

Method	MAE	PCL-10%
MFP	0.54	56.90%
STL-R	0.72	58.38%
MTL-DR	0.68	64.45%
MTCL-Net	0.46	78.93%

3.2.2. Influence of Different Training Data Sets on Range Estimation Performance

Further research was conducted on the localization performance of the MTCL-Net, MTL-DR, and STL methods using different datasets. Based on the standard dataset without introducing any environmental perturbations, models were trained using the MTCL-Net, MTL-DR, and STL methods respectively, and their performance was verified on the test dataset. Similar to the above steps, based on the environmental perturbed dataset, models were trained using the MTCL-Net, MTL-DR, and STL methods respectively, and performance verification was carried out on an independent test dataset. For visualization, range estimates and their corresponding true values under water depths of 25 m, 75 m, 125 m, and 175 m were selected from the independent test dataset, with the results presented in Figure 6.

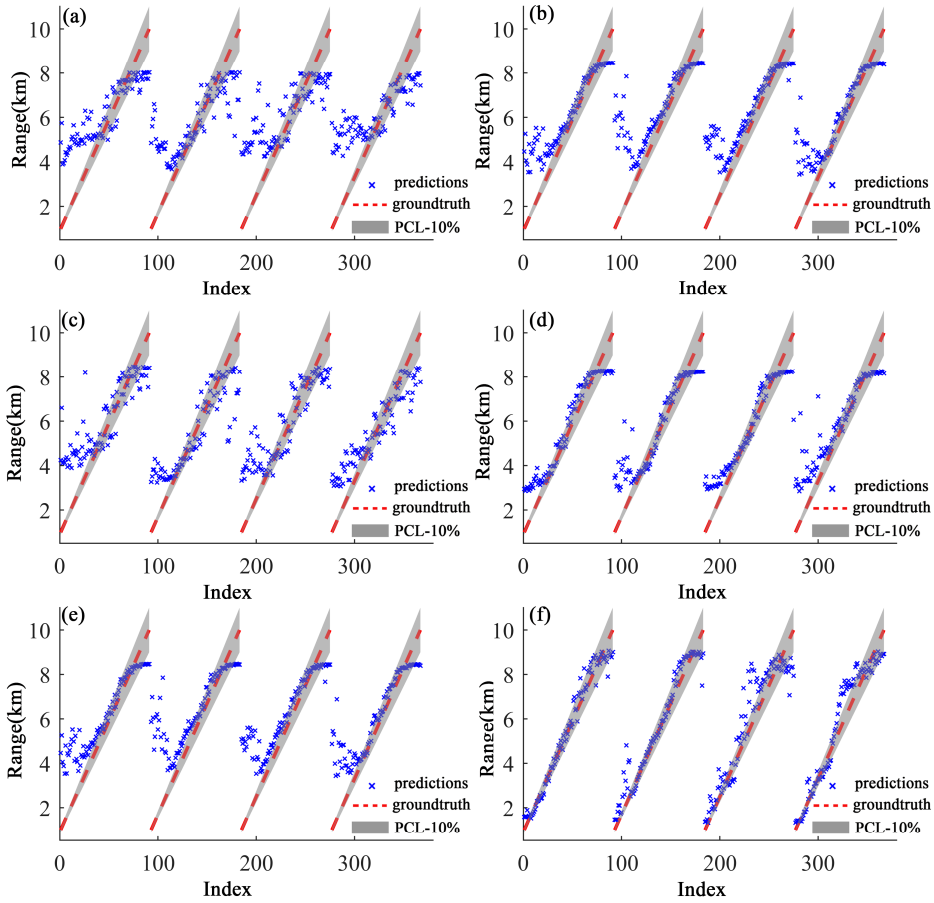


Figure 6. The result of range estimates in test dataset with different methods and different training dataset. (**a**) The range estimates of STL based on standard dataset (STL-S); (**b**) the range estimates of STL based on perturbed dataset (STL-P); (**c**) the range estimates of MTL-DR based on standard dataset (MTL-DR-S); (**d**) the range estimates of MTL-DR based on perturbed dataset (MTL-DR-P); (**e**) the range estimates of MTCL-Net based on standard dataset (MTCL-S); (**f**) the range estimates of MTCL-Net based on perturbed dataset (MTCL-P).

Separate training was conducted based on the standard dataset and the perturbed dataset, and the visualization of the prediction results of the single-task learning model compared with the independent test set is shown in Figure 6a,b. The range prediction results of the STL trained on the standard dataset exhibit a "discrete" state, with most prediction results deviating from the PCL-10% error range. In contrast, the range prediction results of the STL method trained on the perturbed dataset are closer to the trend of the true range labels. Although the errors of some samples deviate from the PCL-10% error range, there is a significant improvement compared with the prediction results based on the standard dataset. The same results are also reflected in the MTL-DR method (as shown in Figure 6c,d) and the MTCL-Net method (as shown in Figure 6e,f). The models trained on the perturbed dataset show a significantly better distance prediction performance than those trained on the standard dataset.

Meanwhile, it is found that deep learning-based methods exhibit significant errors at both the initial and final ranges in distance estimation. The training and test datasets in this study are constructed within a finite ranging interval. For samples located at the two ends of the distance interval (i.e., the initial near-distance stage and the final far-distance stage, exactly corresponding to the deviation regions observed in Figures 5 and 6), their input acoustic pressure features only have one-sided adjacent sample support in the distance dimension. In contrast, samples in the middle of the interval can obtain sufficient feature reference and learning constraints from bilateral adjacent samples in the distance dimension. During model training, the feature patterns of middle-interval samples present a more abundant and stable statistical distribution, enabling the model to learn a more robust nonlinear mapping between acoustic features and distance labels. However, the feature patterns of boundary samples lack sufficient adjacent data support, leading to poor generalization of the trained model on these edge samples.

The MAE and PCL-10% results of the models trained on the standard dataset and the perturbed dataset on the test dataset are shown in Table 3. The models trained on the perturbed dataset outperform those trained on the standard dataset in the prediction results on the test dataset, indicating that constructing a perturbed dataset by introducing random perturbations can enhance the model's prediction performance. Introducing appropriate noise during model training is an effective regularization strategy, which can significantly improve the model's generalization ability and robustness. By applying controllable random perturbations to the input data or the internal state of the model, the model is forced to learn more universal structures in the data distribution, rather than accidental deviations or noise in the training samples. Secondly, the random combination of environmental parameters can enhance the model's robustness to input uncertainty. Simulating measurement errors, sensor noise, or distribution shifts in real environments during the training phase can enable the model to maintain prediction stability when facing perturbations. In addition, noise can introduce random gradient fluctuations in the optimization process, which helps parameter updates escape local optima and promotes the model to converge to a smoother minimum region with a better generalization performance. In deep neural networks, noise can also improve the model's feature representation ability. For example, through mechanisms such as dropout or random batch normalization, the model can learn diverse feature representations under different perturbed conditions, thereby enhancing the robustness and expressiveness of features. In general, noise injection effectively limits the model complexity through implicit regularization, guiding it to learn a more stable and generalizable solution.

Table 3. Performance comparison of different methods based on different datasets.

Method	MAE	PCL-10%
STL-S	1.38 km	33.30%
STL-P	1.04 km	45.91%
MTL-DR-S	1.06 km	40.55%
MTL-DR-P	0.81 km	54.48%
MTCL-S	1.30 km	30.20%
MTCL-P	0.54 km	65.16%

3.2.3. Effect of Spatial Constraints on the Performance of Range Estimation

In order to further evaluate the effectiveness of the spatial similarity constraint structure proposed in this study, the t-SNE method is used to visually display the deep-seated features extracted from the network structure [27]. By mapping high-dimensional data to two-dimensional space, the t-SNE method can measure the probability distribution of high-dimensional data under the condition of ensuring the consistency of data similarity before and after mapping.

The data similarity of high-dimensional space can be expressed as

$$\mathrm{Sim}_{i|j}^{\mathrm{before}} = \frac{\exp\left(\frac{-\|a_i - a_j\|^2}{2\sigma_i^2}\right)}{\Sigma_{k\neq i}\exp\left(\frac{-\|a_i - a_k\|^2}{2\sigma_i^2}\right)}, \tag{12}$$

where *ai* represents the samples in high space, ||~|| is Euclidean distance, σi represents the superparameter of samples, and the data similarity of two-dimensional space can be expressed as

$$\mathrm{Sim}_{i|j}^{\mathrm{after}} = \frac{\left(1 + \|b_i - b_j\|^2\right)^{-1}}{\Sigma_{k\neq i}\left(1 + \|b_i - b_k\|^2\right)^{-1}}, \tag{13}$$

where *bi* is the samples in two-dimensional space. To ensure the similarity in the new two-dimensional space, the optimization objective can be expressed as

$$\mathrm{Objective} = \sum_i \sum_j \mathrm{Sim}_{i|j}^{\mathrm{before}} \log\left(\frac{\mathrm{Sim}_{i|j}^{\mathrm{before}}}{\mathrm{Sim}_{i|j}^{\mathrm{after}}}\right). \tag{14}$$

The sample representation in the two-dimensional space can be obtained by optimizing the objective function. Models were trained using single-task learning, multi-task learning, and spatial constraint-based multi-task learning methods respectively based on the standard dataset, perturbed dataset, and mixed dataset. On the independent test set, the t-SNE method was used to reduce the dimensionality of the 512 features in the fully connected layer of CNN-SPS to the two-dimensional space, and the results are shown in the figures. Overall, the deep features extracted by the models trained on the standard dataset are scattered in the two-dimensional space, and the feature distributions under different labels are intertwined, as shown in Figure 7a,d,g. After visualizing the features extracted by the models trained on the perturbed dataset (Figure 7b,e,h), the sample features with similar labels are obviously concentrated and present an aggregated state. In contrast, the features extracted by the models trained on the mixed dataset show a more obvious separation trend, indicating that training models under the mixed dataset state helps to improve the separability of features. Intuitively, there are also obvious differences in the feature visualization results of models trained based on different methods. As shown in

Figure 7b,e,h, the feature extraction effect of the model based on the MTCL-Net method is superior to that of the models based on the STL and MTL-DR methods, and the same effect is observed in Figure 7c,f,i.

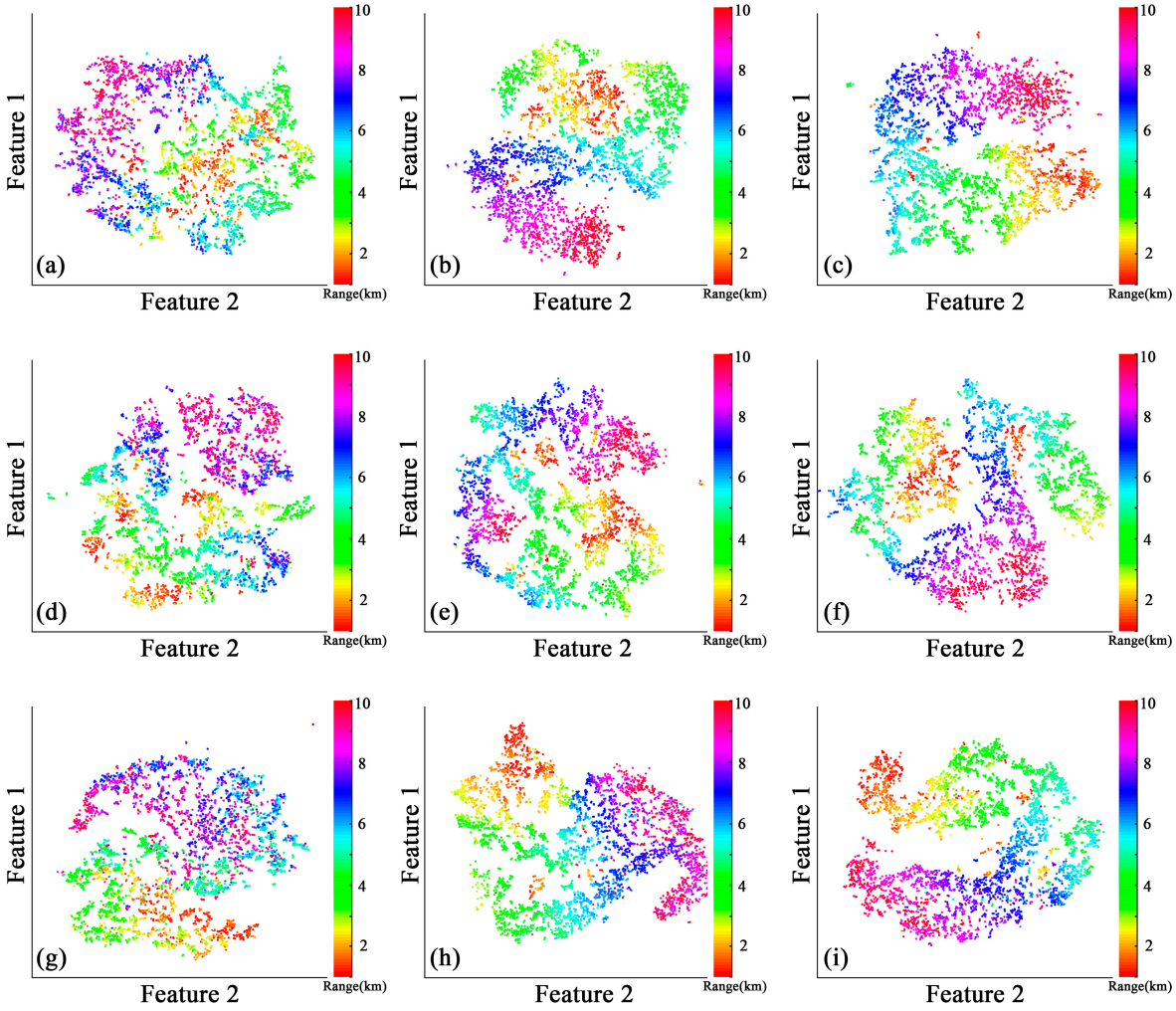


Figure 7. Two-dimensional feature distributions of different methods. (**a**) STL-S; (**b**) STL-P; (**c**) STL method based on mixed dataset; (**d**) MTL-DR-S; (**e**) MTL-DR-P; (**f**) MTL-DR based on mixed dataset; (**g**) MTCL-S; (**h**) MTCL-P; (**i**) MTCL method based on mixed dataset.

To more effectively evaluate the feature extraction performance under different datasets and methods, this study calculated the ratio of the average intra-class distance to the average inter-class distance for quantitative evaluation and comparison [43]. This ratio can be used to measure the clustering effect: a smaller ratio indicates that samples of the same class are more compactly distributed, samples of different classes are more distinctly separated, the quality of clustering results is higher, the feature separability is greater, and it is more conducive to the discrimination of final results. The ratios of the average intra-class distance to the average inter-class distance of feature extraction results under different datasets and methods are shown in Figure 8. It can be seen that the MTCL-Net method based on the mixed dataset has the smallest ratio, indicating the best feature separation effect. This demonstrates that the MTCL-Net method established in this study has an excellent feature extraction capability, which can effectively distinguish acoustic source features under different distance labels and facilitate subsequent acoustic source localization.

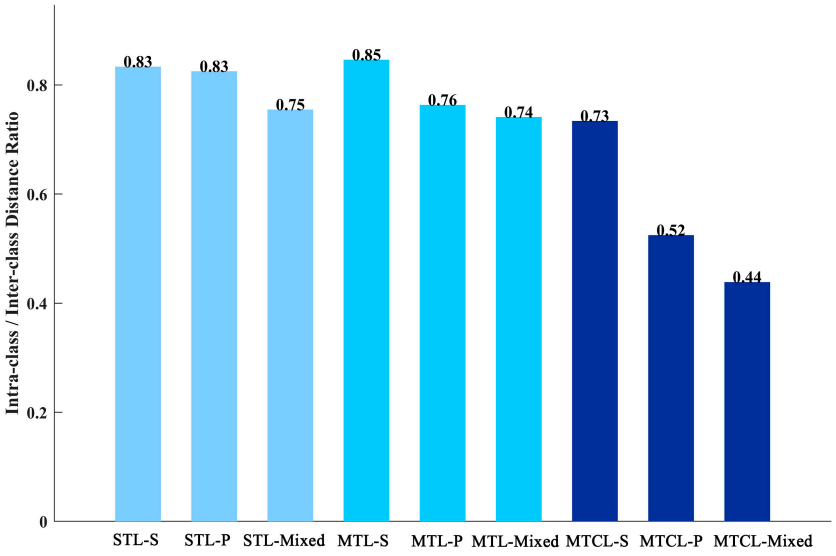


Figure 8. The feature extraction performance under different methods.

Further analysis was conducted on the features extracted by the CNN-SPS structure under different training iterations. The feature similarity discrimination based on the Siamese network is a constrained task for acoustic source distance estimation, which helps to extract features related to the acoustic source position and thereby affects the distance estimation performance of the model. This section investigates the feature extraction performance of the CNN-SPS under different training iterations. According to the experimental conditions, the number of iterations was set to range from 6000 to 36,000, and the feature extraction results under different training iterations were visualized, with the results shown in Figure 9. In the figure, Feature1 and Feature2 are directly output by the two neurons at the end of the feature training network, and the continuous color band represents the distance label value (from 1 km to 11 km). Figure 9 reflects the separability of the acoustic source distance features extracted by the overall network structure through training in the two-dimensional space. With the continuous increase in the number of iterations, the samples originally clustered together are gradually separated. Since the labels corresponding to different samples are the distances between the acoustic source and the array, the final two-dimensional feature visualization results are similar to a "linear shape", mapping samples of different distances. This indicates that the feature extraction network established in this study can acquire features related to the acoustic source distance, and with the increase in training iterations, the feature separability continues to increase, which helps to form constraints on the acoustic source distance estimation task and improve the model performance.

The MAE and PCL of the MTCL-Net method under different iterations are as shown in Figure 10. With the increase in iteration times, the performance of the model gradually improves. After about 30,000 iterations, the MAE and PCL-10 indexes of the model tend to remain unchanged, and the performance of the model is no longer improved.

In the MTCL-Net method, constructing sample pairs in the standard dataset and the perturbed dataset is a prerequisite for spatial constraint. Among them, the Δd and Δr parameters represent the spatial consistency search distances in the depth and range directions, respectively. Their selection directly affects the subsequent model training effect. To pursue the optimal model performance, this study conducted comparative experiments based on different depth and distance direction parameters, and verified the performance on the independent test set using MAE and PCL as indicators, with the results shown in the Figure 11. When Δr = 30 m and Δd = 25 m, that is, selecting the range of 30 m in the distance direction and 25 m in the depth direction as the constrained

space, the trained model achieves the optimal effect. All experiments involved in this study are obtained under this parameter by default. It should be noted that the selection of these parameters is related to the grid size setting of the distance and depth method, environmental disturbance parameters, etc. At present, it is not possible to accurately give the selection criteria and basis, and it can only be determined by comparing a large number of parameter selection experiments.

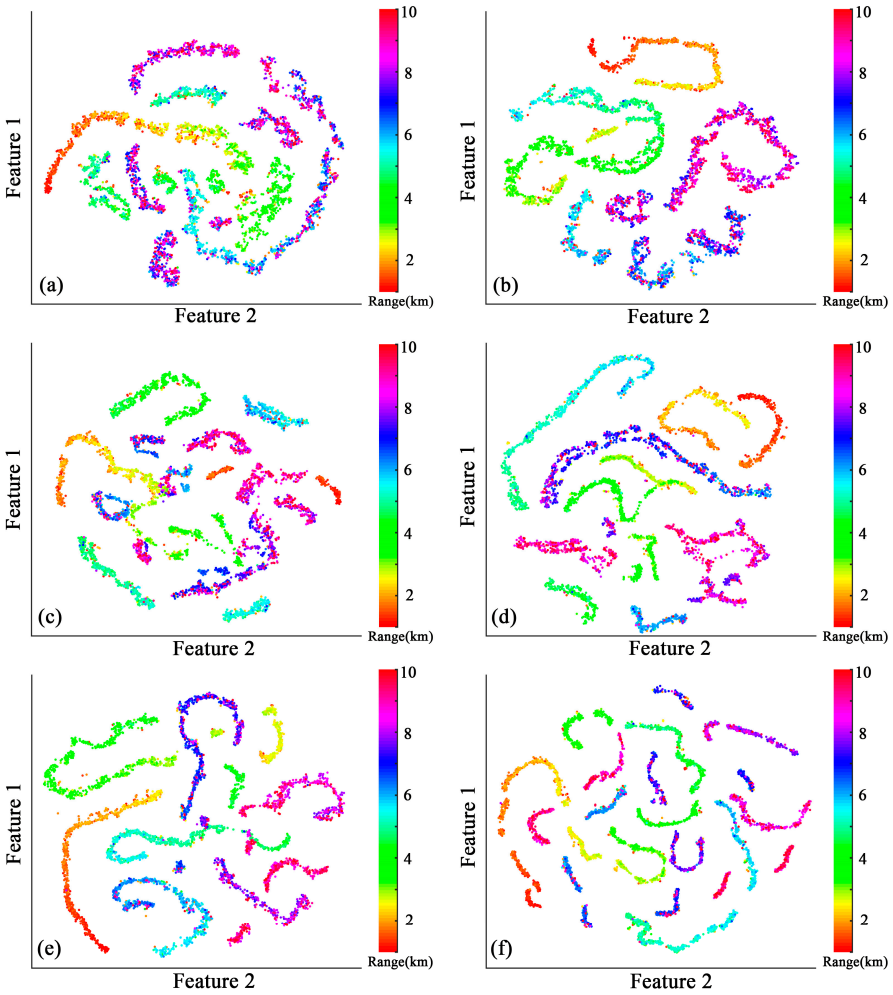


Figure 9. Visualization of features extracted by MTCL-Net at different training iterations. (**a**) 6000 iterations; (**b**) 12,000 iterations; (**c**) 18,000 iterations; (**d**) 24,000 iterations; (**e**) 30,000 iterations; (**f**) 36,000 iterations.

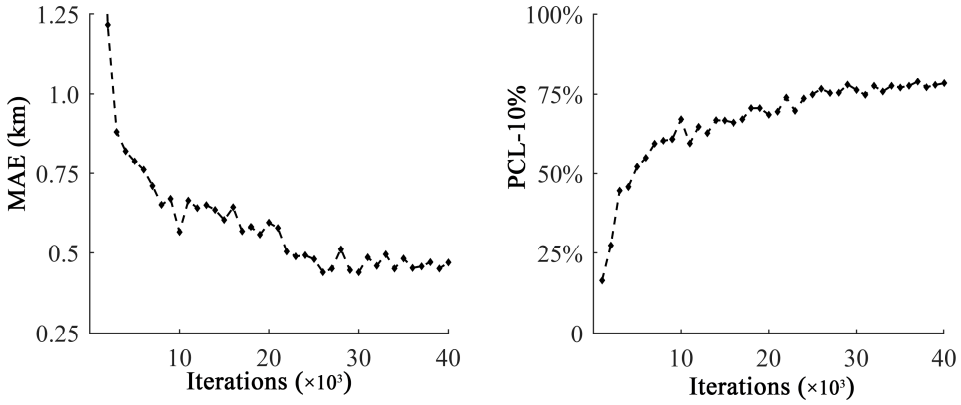


Figure 10. The MAE and PCL-10 of MTCL-Net method under different iterations.

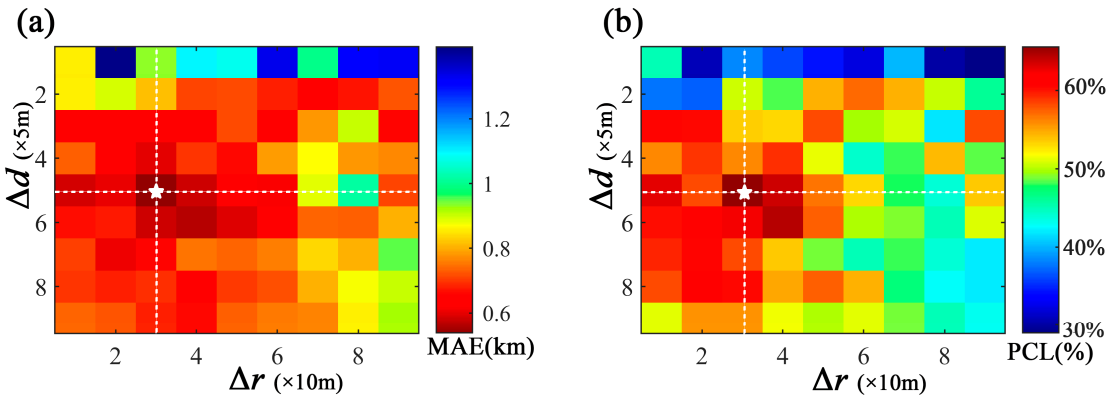


Figure 11. Influence of different range parameters on model performance, where the dotted line and star represent the Optimum parameters. (**a**) The MAE; (**b**) the PCL-10%.

3.3. Sea Trial Experiments and Results

3.3.1. Overview of SWellEx-96 Experiment

The sea trial data used in this paper comes from the SWellEx-96 experiment conducted in May 1996 in the waters near San Diego, California. During the S5 leg of the experiment (as shown in Figure 12) there was a deep source (approximately 54 m) and a shallow source (approximately 9 m) while traveling from south to north. The two sources transmitted CW signals at multiple frequencies in the ranges of 49–400 Hz and 109–385 Hz, respectively, with no overlapping frequency points. A vertical line array (VLA) with 21 elements was employed to record the acoustic responses for underwater source localization, with a sampling frequency of 1500 Hz, and a total of 75 min of data were acquired. This paper mainly processes the signals transmitted by the deep source, using the same frequency set as in the simulation experiments: {49, 64, 79, 94, 112, 130, 148, 166, 201, 235, 283, 338} Hz.

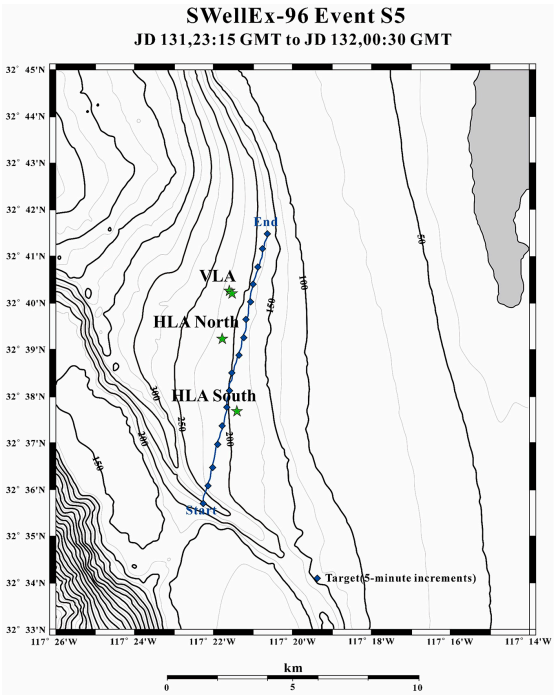


Figure 12. Schematic of the SWellEx-96 experiment S5 voyage.

According to the official documentation, LFM signals were transmitted at the beginning, middle, and end of the sea trial, during which no CW signals were emitted. In actual data processing, multiple LFM signal transmissions were observed. The corresponding

time segments were removed to avoid interference from invalid data. Finally, four continuous time intervals were retained for processing: 1–1021 s, 1321–2280 s, 2341–3300 s, and 3600–4500 s.

3.3.2. Range Estimation Results and Analysis

Range estimation is performed on the SWellEx-96 sea trial measured data using four compared methods, namely MFP, STL, MTL-DR, and the proposed MTCL-Net. During data preprocessing, the data are segmented with a step size of 2 s and an overlap ratio of 50%. A normalized kaiser window is applied to each segmented signal, followed by Fourier transform to obtain the acoustic pressure. The ± 1 frequency bins adjacent to the target frequency are searched, and the FFT value corresponding to the maximum power is extracted to accommodate the Doppler shift. Finally, the measurement matrix is acquired and normalized. A total of 3840 valid samples are constructed from the collected sea trial measurements. For the data-driven deep learning models (STL, MTL-DR, and MTCL-Net), a pre-training and fine-tuning strategy is adopted to address the limitation of insufficient measured samples: all models are pre-trained on the simulated datasets and then fine-tuned on a subset of the measured data to validate their range estimation performance in real marine environments. Following the equal-interval sampling principle, 10% of the total measured samples (384 samples) are selected as the independent held-out test dataset, with the remaining 90% of samples used for model fine-tuning. All deep learning-based methods employ ResNet-18 as the feature extraction architecture, followed by a CNN network for further distance estimation. In this study, the weights of the ResNet-18 architecture are therefore frozen and treated solely as a feature extractor, while only the CNN network undergoes parameter fine-tuning. During fine-tuning, the learning rate is set to 0.001 and the learning rate decay factor to 0.9. Iteration is stopped when the test accuracy no longer improves. The performance of all fine-tuned deep learning models, along with the physics-based MFP method, is systematically evaluated and compared on this independent test dataset.

The range estimation results of the conventional MFP method on the independent test dataset are presented in Figure 13. MFP achieves a relatively high ranging accuracy in the first and last time intervals of the sea trial, where the relative errors between the estimated ranges and the ground truth values mostly fall within the 10% error bound. However, significant deviations occur in the range estimation results from approximately the 25th minute to the 55th minute of the experiment, with relative errors exceeding the 10% credible localization threshold defined in this work. This performance degradation is mainly attributed to the mismatch between the assumed nominal environmental parameters and the time-varying real marine environment during the sea trial, which directly reflects the inherent limitation of the MFP method, that is prone to severe localization performance degradation in the presence of marine environmental parameter mismatches.

The range estimation results of the STL method on the independent test dataset are shown in Figure 14. The estimation accuracy is relatively high during the first half of the experiment, and the errors relative to the true ranges mostly fall within the 10% error bound. However, large deviations appear in the range estimation results during the second half of the experiment, and the error is particularly severe at the closest point of the source, far exceeding the 10% error bound. This may be related to the simulation dataset used in the model pre-training stage. The range labels of the simulation dataset are distributed from 1 km to 10 km, and the model tends to focus on features corresponding to the middle ranges during iterative training. As a result, the model exhibits a poor feature learning ability near the distance boundaries, leading to large range estimation deviations at the beginning of the experiment and at the closest position to the source.

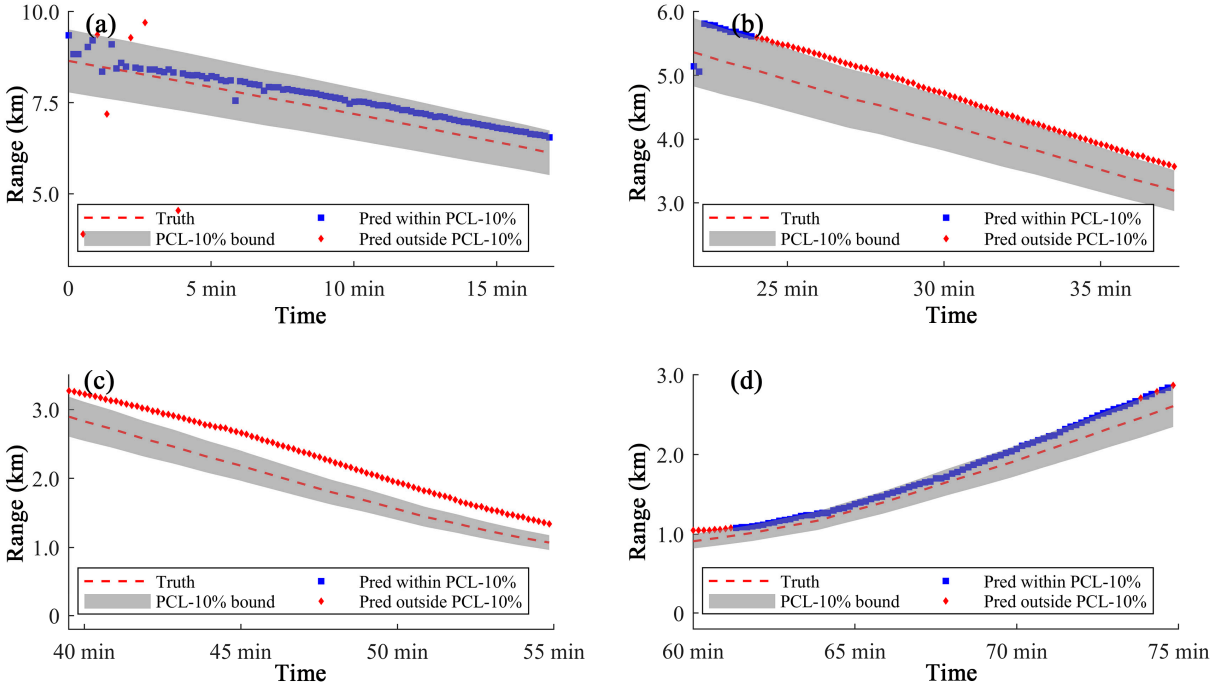


Figure 13. The range estimation result of MFP method. (**a**) 1–1021 s; (**b**) 1321–2280 s; (**c**) 2341–3300 s; (**d**) 3600–4500 s.

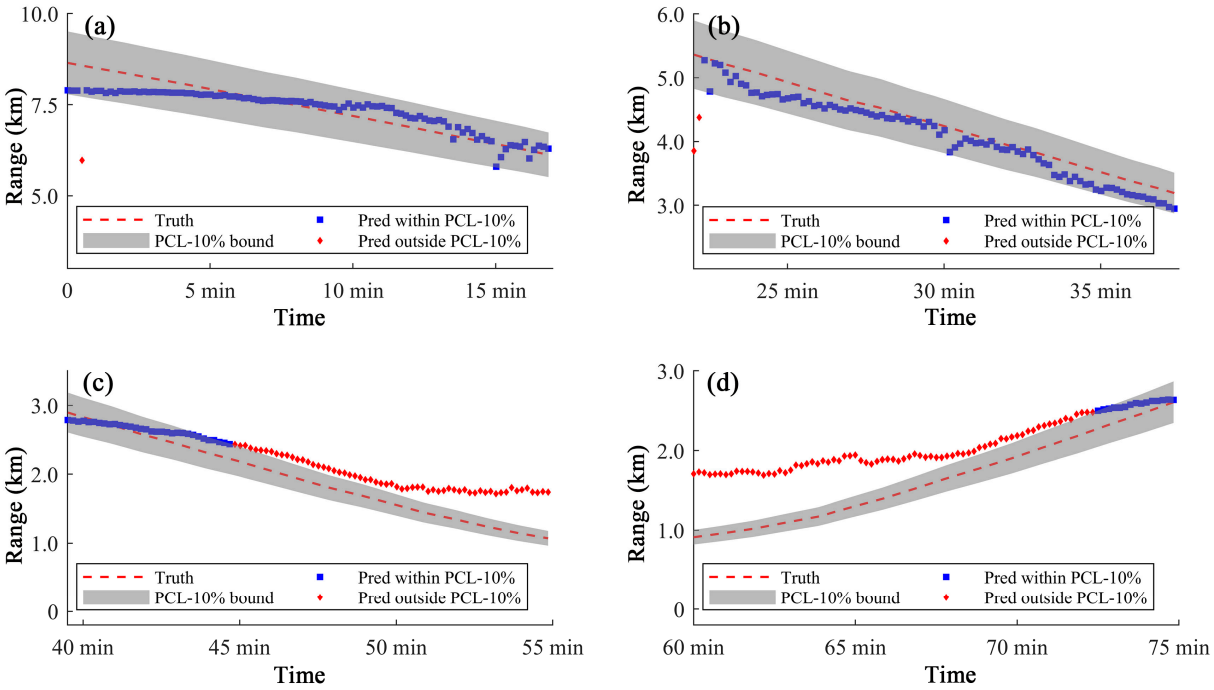


Figure 14. The range estimation result of STL method. (**a**) 1–1021 s; (**b**) 1321–2280 s; (**c**) 2341–3300 s; (**d**) 3600–4500 s.

The range estimation results of the MTL-DR method on the independent test dataset are shown in Figure 15. Similar to the STL method, the estimation accuracy is relatively high during the first half of the experiment, with most errors relative to the true ranges falling within the 10% error bound. Although large deviations still appear at the closest position to the source during the experiment, the performance is significantly improved compared with the single-task learning method, and the range estimation results are evidently superior to those of the STL method.

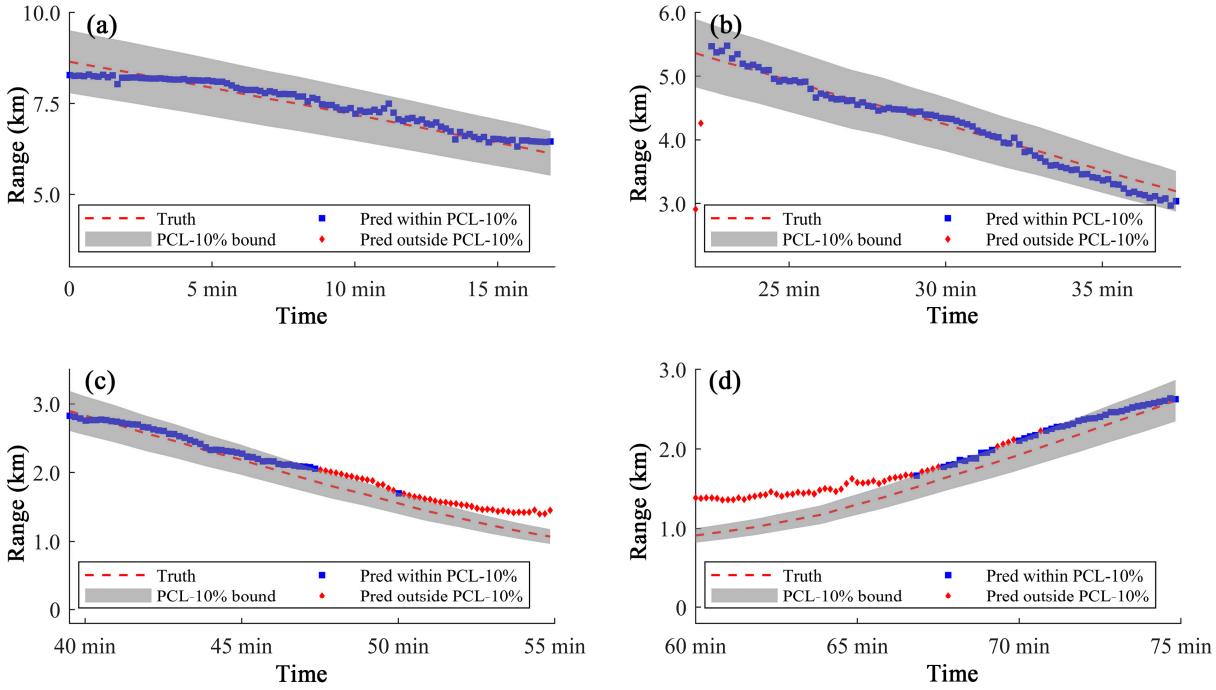


Figure 15. The range estimation result of MTL-DR method. (**a**) 1–1021 s; (**b**) 1321–2280 s; (**c**) 2341–3300 s; (**d**) 3600–4500 s.

The range estimation results based on the MTCL-Net method are shown in Figure 16. Although the estimation error is relatively large near the start time, it basically remains within the 10% error bound of the true range. A certain estimation error also exists at the moment when the source is closest to the receiver, which is attributed to similar reasons as in the STL and MTL-DR methods. Nevertheless, the MTCL-Net method achieves the best range estimation performance among all the compared approaches.

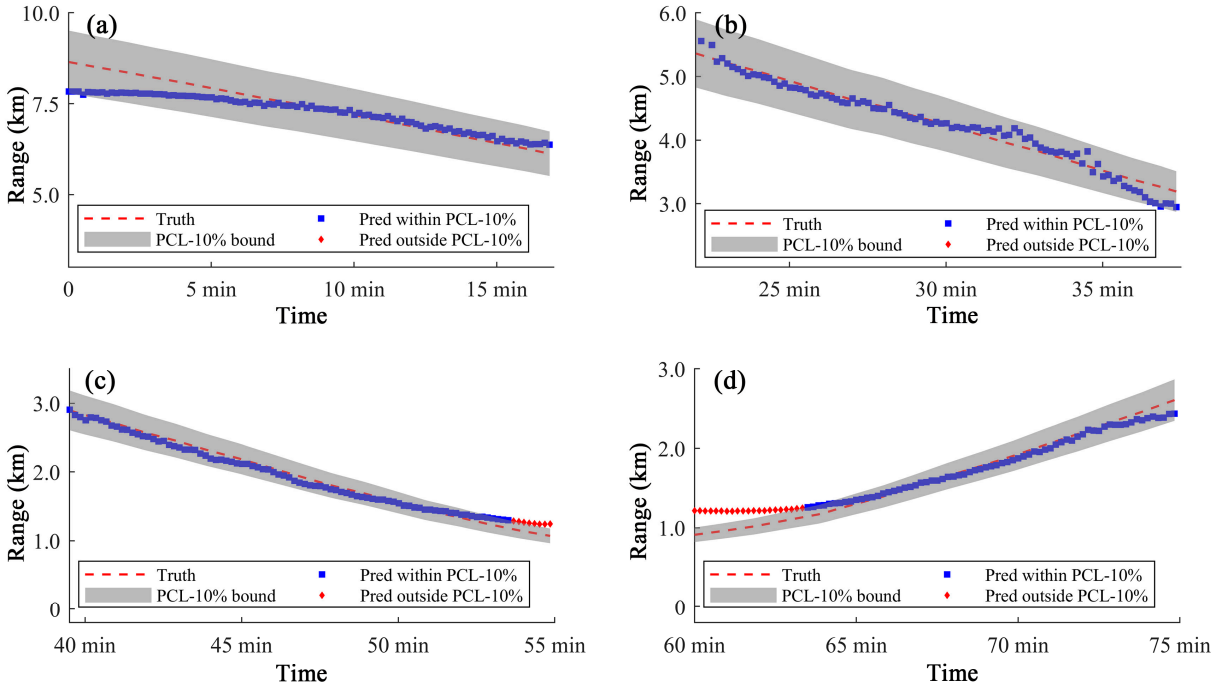


Figure 16. The range estimation result of MTCL-Net method. (**a**) 1–1021 s; (**b**) 1321–2280 s; (**c**) 2341–3300 s; (**d**) 3600–4500 s.

The MAE and PCL metrics of different methods are listed in Table 4. Consistent with the above results, the MTCL-Net method achieves the best ranging performance among all methods, with the smallest MAE of 0.17 km and the highest PCL-10% of 90.36%. It outperforms the traditional MFP method and STL. Meanwhile, the proposed spatial constraint mechanism also shows clear advantages over the conventional MTL-DR method, leading in both MAE and PCL-10% metrics.

Table 4. Range estimation performance of different methods.

Methods	MAE	PCL-10%
MFP	0.45 km	48.70%
STL	0.38 km	62.24%
MTL-DR	0.18 km	72.66%
MTCL-Net	0.17 km	90.36%

To further evaluate the influence of the number of samples used for model fine-tuning on the final performance, this study selected different numbers of samples from the measured data for model fine-tuning. The MAE and PCL-10% metrics of the range estimation results on the independent test dataset were compared and analyzed. The proportion of samples used for fine-tuning ranged from 10% to 90%, and the results are shown in Figure 17.

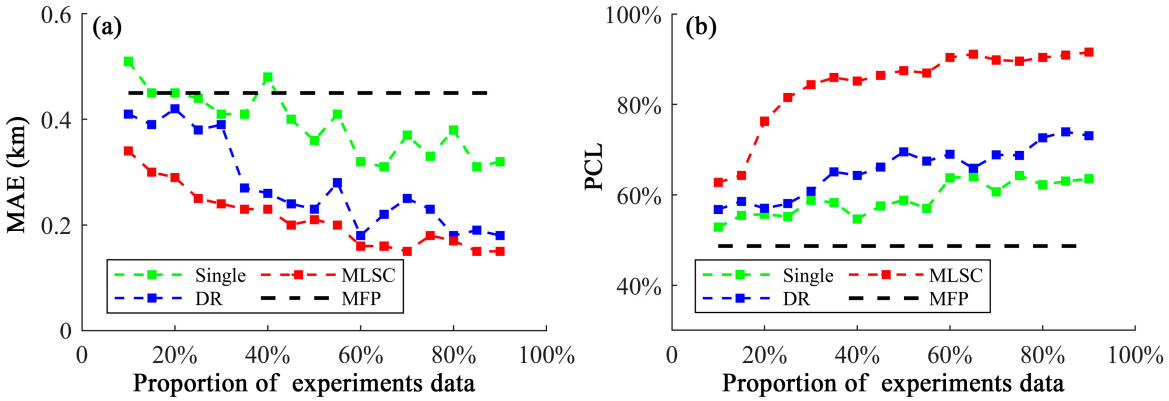


Figure 17. Model performance under different numbers of fine-tuning samples. (**a**) The MAE; (**b**) The PCL-10%.

As the number of fine-tuning samples increases, the range estimation accuracy of all deep learning methods improves continuously. Under different sample sizes, the MTCL-Net method maintains the best performance, with the smallest range estimation error and the highest PCL-10% value, indicating that its ability to extract and represent distance features is superior to other methods. It can be observed from the figure that in this experiment, the MTCL-Net method achieves an optimal performance using only about 60% of the samples for fine-tuning. In contrast, the single-task learning and distance–depth multi-task learning methods show a continuously decreasing MAE and increasing PCL-10% as the sample size grows. Therefore, compared with the traditional single-task learning and distance–depth-based multi-task learning methods, the MTCL-Net method requires fewer measured samples for fine-tuning and exhibits greater advantages in practical applications.

4. Conclusions

To address the poor performance of underwater acoustic source ranging models in practical applications, which arises from uncertain marine environmental parameters and limited measured underwater acoustic data, this paper proposes a multi-task learning

network that incorporates contrastive learning as an auxiliary task (MTCL-Net). In this method, the perturbed dataset was constructed using a random combination of key marine environmental parameters, which can significantly enhance the model's generalization capability and environmental robustness by forcing it to learn universally distributed structural features and mitigate overfitting. In addition, the Siamese network module based on contrastive learning for acoustic source spatial position similarity discrimination can guide the model to learn discriminative position-related features, notably boosting its feature extraction capability, as verified by t-SNE visualization and intra-class/inter-class distance quantitative evaluation; the multi-task learning mechanism integrating standard dataset localization, perturbed dataset localization and position similarity discrimination tasks establishes effective inter-task mutual constraints via the shared ResNet-18 feature extraction backbone, enabling more accurate and stable feature representation and a superior localization performance over traditional single-task learning and distance–depth multi-task learning methods; finally, the proposed method exhibits excellent practicality in real marine environments, achieving an MAE of 0.17 km and a PCL-10% of 90.36% on SWellEx-96 measured data to outperform all comparative methods, while only requiring about 60% of measured samples for fine-tuning to reach an optimal performance, effectively alleviating the practical challenge of scarce labeled underwater acoustic data.

In conclusion, the MTCL-Net method proposed in this paper enhanced the accuracy, robustness, and practicality of the ranging model. It provides a novel technical approach and methodological support for underwater acoustic source localization in complex marine environments. Future research may further optimize the parameter configuration strategy for the spatial neighborhood range in the Siamese network, explore more efficient multi-task loss function allocation methods, and extend the model's application scope to more complex marine environments and various types of acoustic sources. In addition to that, the current method is validated and optimized for the range-independent shallow water waveguide environment. For ocean scenarios with severe range-dependent environmental changes (such as areas with frequent internal waves, ocean fronts, or complex seabed topography), the model's ability to suppress environmental interference will be reduced to a certain extent, and the performance of the model remains to be further evaluated.

Author Contributions: Conceptualization, J.Z., B.M. and Z.Q.; methodology, J.Z.; software, B.M.; validation, J.Z.; formal analysis, J.Z.; investigation, J.Z.; resources, Z.Q.; data curation, B.M.; writing—original draft preparation, J.Z. and B.L.; writing—review and editing, J.Z. and S.P.; visualization, J.Z.; supervision, Z.Q.; project administration, Z.Q. and W.L.; funding acquisition, B.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Taishan Scholars Program, Shandong Provincial Natural Science Foundation (grant number ZR2024QD082) and the Fundamental Research Funds for the Central Universities under grant no. 3072025KX2601.

Data Availability Statement: Publicly available datasets were analyzed in this study. The original data presented in the study are openly available at http://swellex96.ucsd.edu/index.htm (accessed on 12 December 2025).

Acknowledgments: Thanks to all participants in the SwellEX-96 experiment for their selfless data contribution. We express our sincere gratitude for the financial support from the Taishan Scholars Program, Shandong Provincial Natural Science Foundation (grant number ZR2024QD082). Their generous funding has made it possible for us to conduct this research and has significantly contributed to the success of our work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Baggeroer, A.B.; Kuperman, W.A.; Mikhalevsky, P.N. An Overview of Matched Field Methods in Ocean Acoustics. *IEEE J. Ocean. Eng.* **1993**, *18*, 401–424. [CrossRef]
2. Sazontov, A.G.; Malekhanov, A.I. Matched Field Signal Processing in Underwater Sound Channels (Review). *Acoust. Phys.* **2015**, *61*, 213–230. [CrossRef]
3. Kim, K.; Seong, W.; Lee, C. Matched Field Processing: Analysis of Feature Extraction Method with Ocean Experimental Data. In Proceedings of the 2004 International Symposium on Underwater Technology (IEEE Cat. No.04EX869), Taipei, Taiwan, 20–23 April 2004; pp. 181–185.
4. Feuillade, C.; Del Balzo, D.R.; Rowe, M.M. Environmental Mismatch in Shallow-Water Matched-Field Processing: Geoacoustic Parameter Variability. *J. Acoust. Soc. Am.* **1989**, *85*, 2354–2364. [CrossRef]
5. Hamson, R.M.; Heitmeyer, R.M. Environmental and System Effects on Source Localization in Shallow Water by the Matched-Field Processing of a Vertical Array. *J. Acoust. Soc. Am.* **1989**, *86*, 1950–1959. [CrossRef]
6. Shang, E.C.; Wang, Y.Y. Environmental Mismatching Effects on Source Localization Processing in Mode Space. *J. Acoust. Soc. Am.* **1989**, *86*, S7–S8. [CrossRef]
7. Li, X.; Xiang, L.; Zeng, Y.; Jia, H. Analysis of Mode Mismatch in Uncertain Shallow Ocean Environment. *Appl. Acoust.* **2018**, *139*, 149–155. [CrossRef]
8. Grumiaux, P.-A.; Kitić, S.; Girin, L.; Guérin, A. A Survey of Sound Source Localization with Deep Learning Methods. *J. Acoust. Soc. Am.* **2022**, *152*, 107–151. [CrossRef] [PubMed]
9. Wang, Y.; Peng, H. Underwater Acoustic Source Localization Using Generalized Regression Neural Network. *J. Acoust. Soc. Am.* **2018**, *143*, 2321–2331. [CrossRef]
10. Niu, H.; Reeves, E.; Gerstoft, P. Source Localization in an Ocean Waveguide Using Supervised Machine Learning. *J. Acoust. Soc. Am.* **2017**, *142*, 1176–1188. [CrossRef]
11. Yang, Z.; Zhu, Z.; Zhao, Y.; Tian, Y.; Fan, C.; Guo, R.; Lu, W.; Ge, J.; Chen, B.; Zhang, Y.; et al. A Comprehensive Survey on Underwater Acoustic Target Positioning and Tracking: Progress, Challenges, and Perspectives. *arXiv* **2025**, arXiv:2506.14165. [CrossRef]
12. Niu, H.; Gong, Z.; Ozanich, E.; Gerstoft, P.; Wang, H.; Li, Z. Deep-Learning Source Localization Using Multi-Frequency Magnitude-Only Data. *J. Acoust. Soc. Am.* **2019**, *146*, 211–222. [CrossRef]
13. Liu, M.; Niu, H.; Li, Z.; Guo, Y. Source Depth Estimation with Feature Matching Using Convolutional Neural Networks in Shallow Water. *J. Acoust. Soc. Am.* **2024**, *155*, 1119–1134. [CrossRef] [PubMed]
14. Wang, W.; Wang, Z.; Su, L.; Hu, T.; Ren, Q.; Gerstoft, P.; Ma, L. Source Depth Estimation Using Spectral Transformations and Convolutional Neural Network in a Deep-Sea Environment. *J. Acoust. Soc. Am.* **2020**, *148*, 3633–3644. [CrossRef] [PubMed]
15. Long, R. Range Estimation in a Weakly Range-Dependent Waveguide in the South China Sea Using Single Hydrophone and Multi-Task Trained Deep Neural Network. *Appl. Acoust.* **2023**, *213*, 109630. [CrossRef]
16. Liu, M.; Niu, H.; Li, Z. Implementation of Bartlett Matched-Field Processing Using Interpretable Complex Convolutional Neural Network. *JASA Express Lett.* **2023**, *3*, 026003. [CrossRef]
17. Liu, Y.; Niu, H.; Li, Z. A Multi-Task Learning Convolutional Neural Network for Source Localization in Deep Ocean. *J. Acoust. Soc. Am.* **2020**, *148*, 873–883. [CrossRef]
18. Liu, M.; Niu, H.; Li, Z.; Guo, Y.; Liu, Y.; Liu, J.; Wu, S.; Nie, L. A Convolutional Neural Network Combining Classification and Regression for Source Localization in Shallow Water. *J. Phys. Conf. Ser.* **2023**, *2486*, 012068. [CrossRef]
19. Wang, W.; Ni, H.; Su, L.; Hu, T.; Ren, Q.; Gerstoft, P.; Ma, L. Deep Transfer Learning for Source Ranging: Deep-Sea Experiment Results. *J. Acoust. Soc. Am.* **2019**, *146*, EL317–EL322. [CrossRef]
20. Liu, Y.; Niu, H.; Li, Z.; Wang, M. Deep-Learning Source Localization Using Autocorrelation Functions from a Single Hydrophone in Deep Ocean. *JASA Express Lett.* **2021**, *1*, 036002. [CrossRef]
21. He, J.; Zhang, B.; Liu, P.; Li, X.; Wang, L.; Tang, R. Effective Underwater Acoustic Target Passive Localization of Using a Multi-Task Learning Model with Attention Mechanism: Analysis and Comparison under Real Sea Trial Datasets. *Appl. Ocean Res.* **2024**, *150*, 104072. [CrossRef]
22. Zhu, X.; Dong, H.; Salvo Rossi, P.; Landrø, M. Feature Selection Based on Principal Component Regression for Underwater Source Localization by Deep Learning. *Remote Sens.* **2021**, *13*, 1486. [CrossRef]
23. Jin, P.; Wang, B.; Li, L.; Chao, P.; Xie, F. Semi-Supervised Underwater Acoustic Source Localization Based on Residual Convolutional Autoencoder. *EURASIP J. Adv. Signal Process.* **2022**, *2022*, 107. [CrossRef]
24. Zhu, X.; Dong, H.; Rossi, P.S.; Landro, M. Time-Frequency Fused Underwater Acoustic Source Localization Based on Contrastive Predictive Coding. *IEEE Sens. J.* **2022**, *22*, 13299–13308. [CrossRef]
25. Wen, H.; Yang, C.; Dou, D.; Xu, L.; Jiao, Y. Underwater Source Ranging by Siamese Network Aided Semi-Supervised Learning. *JASA Express Lett.* **2023**, *3*, 094803. [CrossRef] [PubMed]

26. Guo, T.; Song, Y.; Kong, Z.; Lim, E.; Lopez-Benitez, M.; Ma, F.; Yu, L. Underwater Target Detection and Localization with Feature Map and CNN-Based Classification. In Proceedings of the 2022 4th International Conference on Advances in Computer Technology, Information Science and Communications (CTISC), Suzhou, China, 22 April 2022; pp. 1–8.
27. Liu, Y.; Niu, H.; Li, Z.; Zhai, D. Unsupervised Domain Adaptation for Source Localization Using Ships of Opportunity with a Deep Vertical Line Array. *IEEE J. Ocean. Eng.* **2024**, *49*, 180–196. [CrossRef]
28. Huang, Z.; Xu, J.; Gong, Z.; Wang, H.; Yan, Y. Source Localization Using Deep Neural Networks in a Shallow Water Environment. *J. Acoust. Soc. Am.* **2018**, *143*, 2922–2932. [CrossRef]
29. Ge, F.-X.; Bai, Y.; Li, M.; Zhu, G.; Yin, J. Label Distribution-Guided Transfer Learning for Underwater Source Localization. *J. Acoust. Soc. Am.* **2022**, *151*, 4140–4149. [CrossRef]
30. Ferguson, E.L. Multitask Convolutional Neural Network for Acoustic Localization of a Transiting Broadband Source Using a Hydrophone Array. *J. Acoust. Soc. Am.* **2021**, *150*, 248–256. [CrossRef]
31. Jin, K.; Xu, J.; Zhang, X.; Lu, C.; Xu, L.; Liu, Y. An Acoustic Tracking Model Based on Deep Learning Using Two Hydrophones and Its Reverberation Transfer Hypothesis, Applied to Whale Tracking. *Front. Mar. Sci.* **2023**, *10*, 1182653. [CrossRef]
32. Yoon, S.; Yang, H.; Seong, W. Deep Learning-Based High-Frequency Source Depth Estimation Using a Single Sensor. *J. Acoust. Soc. Am.* **2021**, *149*, 1454–1465. [CrossRef]
33. Wang, T.; Su, L.; Ren, Q.; Li, H.; Jia, Y.; Ma, L. A Hybrid–Source Ranging Method in Shallow Water Using Modal Dispersion Based on Deep Learning. *J. Mar. Sci. Eng.* **2023**, *11*, 561. [CrossRef]
34. Qian, P. A Feature-Compressed Multi-Task Learning U-Net for Shallow-Water Source Localization in the Presence of Internal Waves. *Appl. Acoust.* **2023**, *211*, 109530. [CrossRef]
35. Van Komen, D.F.; Neilsen, T.B.; Howarth, K.; Knobles, D.P.; Dahl, P.H. Seabed and Range Estimation of Impulsive Time Series Using a Convolutional Neural Network. *J. Acoust. Soc. Am.* **2020**, *147*, EL403–EL408. [CrossRef]
36. Neilsen, T.B.; Escobar-Amado, C.D.; Acree, M.C.; Hodgkiss, W.S.; Van Komen, D.F.; Knobles, D.P.; Badiey, M.; Castro-Correa, J. Learning Location and Seabed Type from a Moving Mid-Frequency Source. *J. Acoust. Soc. Am.* **2021**, *149*, 692–705. [CrossRef]
37. Van Komen, D.F.; Howarth, K.; Neilsen, T.B.; Knobles, D.P.; Dahl, P.H. A CNN for Range and Seabed Estimation on Normalized and Extracted Time-Series Impulses. *IEEE J. Ocean. Eng.* **2022**, *47*, 833–846. [CrossRef]
38. Murray, J.; Ensberg, D. The Swellex-96 Experiment. 1996. Available online: http://swellex96.ucsd.edu/index.htm (accessed on 1 June 2025).
39. Porter, M. The KRAKEN Normal Mode Program. 2018. Available online: http://oalib.hlsresearch.com/AcousticsToolbox/ (accessed on 1 June 2025).
40. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
41. Hadsell, R.; Chopra, S.; LeCun, Y. Dimensionality Reduction by Learning an Invariant Mapping. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition—Volume 2 (CVPR'06)*; IEEE: New York, NY, USA, 2006; Volume 2, pp. 1735–1742.
42. Chen, R.; Schmidt, H. Model-Based Convolutional Neural Network Approach to Underwater Source-Range Estimation. *J. Acoust. Soc. Am.* **2021**, *149*, 405–420. [CrossRef]
43. Tan, S.; Liu, L.; Peng, C.; Shao, L. Image-to-Class Distance Ratio: A Feature Filtering Metric for Image Classification. *Neurocomputing* **2015**, *165*, 211–221. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.