Adaptive Control Attention Network for Underwater Acoustic Localization and Domain Adaptation

Quoc Thinh Vo, Joe Woods, Priontu Chowdhury, David K. Han
Department of Electrical and Computer Engineering, *Drexel University*, Philadelphia, PA, USA
{qv23, jw3897, pc833, dkh42}@drexel.edu

***Abstract*—Localizing acoustic sound sources in the ocean is a challenging task due to the complex and dynamic nature of the environment. Factors such as high background noise, irregular underwater geometries, and varying acoustic properties make accurate localization difficult. To address these obstacles, we propose a multi-branch network architecture designed to accurately predict the distance between a moving acoustic source and a receiver, tested on real-world underwater signal arrays. The network leverages Convolutional Neural Networks (CNNs) for robust spatial feature extraction and integrates Conformers with self-attention mechanism to effectively capture temporal dependencies. Log-mel spectrogram and generalized cross-correlation with phase transform (GCC-PHAT) features are employed as input representations. To further enhance the model performance, we introduce an Adaptive Gain Control (AGC) layer, that adaptively adjusts the amplitude of input features, ensuring consistent energy levels across varying ranges, signal strengths, and noise conditions. We assess the model's generalization capability by training it in one domain and testing it in a different domain, using only a limited amount of data from the test domain for fine-tuning. Our proposed method outperforms state-of-the-art (SOTA) approaches in similar settings, establishing new benchmarks for underwater sound localization.**

***Index Terms*—time-frequency domain, sound source localization, underwater acoustics, adaptive attention mechanism**

I. Introduction

Sound source localization (SSL) is essential in underwater signal processing, with applications in communication, robotics, and surveillance [1], [2]. Varying water sound speed profiles, bottom sediment properties, surface dynamics, and background noise create complex propagation paths and interference, challenging accurate localization.

Traditional SSL methods improve accuracy using physics-based techniques. Matched Field Processing (MFP) [3] uses known environment models to estimate source location, leveraging spatial patterns across sensor arrays for precise localization. It often relies on propagation models like KRAKEN [4] to generate theoretical field replicas. Other traditional techniques include ESPRIT [5], beamforming [6], and Time Difference of Arrival [7]. These methods often struggle in complex acoustic conditions and depend on detailed environmental priors such as sound speed profiles, sediment properties, and bathymetry.

Recent advances in machine learning (ML) have significantly improved the performance of underwater signal and image processing tasks such as SSL, detection, and classification. CNNs have shown strong capabilities in sonar image recognition and underwater object detection, enabling automated identification and optimization tasks in complex underwater environments [8]–[10]. Recurrent Neural Networks (RNNs), including hybrid architectures, have been effectively applied for time-series modeling and anomaly detection in acoustic sensing systems [11], [12]. More recently, Transformer-based models have demonstrated robust performance in underwater acoustic classification and target recognition, even under noisy conditions, by capturing long-range temporal and spectral dependencies [13], [14]. Additional information on the recent state of underwater acoustic signal processing for various tasks, including SSL, can be found in [15].

The SWellEx-96 dataset [16] has been widely adopted as a benchmark for evaluating the performance of underwater SSL methods across a wide spectrum of approaches, from traditional signal processing techniques to deep learning models such as CNNs and Transformers. Notable approaches include Chen et al. [17] with the extension of CNNs enhancing performance in challenging environments and Chi et al. [18] addresses overfitting using fitting-based early stopping (FEAST). Zhu et al. [19], Wen et al. [20], and Zhu et al. [21] propose two-stage semi-supervised and self-supervised approaches to tackle data scarcity. However, their solutions are computationally intensive due to additional training stages and underperform compared to SOTA methods. Moreover, current SOTA methods such as Wang et al. [22] with U-Net and He et al. [23] with MLF-TransCNN, employ single-branch architectures for feature extraction. This design limits the ability to capture the complexity of multiple feature representations during extraction. In contrast, a multi-branch architecture allows independent extraction in each branch while leveraging correlations among distinct feature sets, thereby enhancing learning. Incorporating multiple features enriches input representations, improving accuracy, robustness, and adaptability, particularly in complex and noisy environments.

To address the limitations of existing methods, we propose an Adaptive Control Attention Network (ACA-Net) featuring a multi-branch architecture where the branches share weights, facilitating a soft-sharing of parameters between them. Log-mel spectrogram and GCC-PHAT [24] features are used as inputs, with the former encoding spectral energy and the latter capturing inter-channel time delays. Additionally, we integrate an AGC layer, which dynamically adjusts the amplitude of input features, ensuring consistent energy levels to improve the model's performance and robustness. We also use fine-tuning for domain adaptation demonstrating that the model can adapt

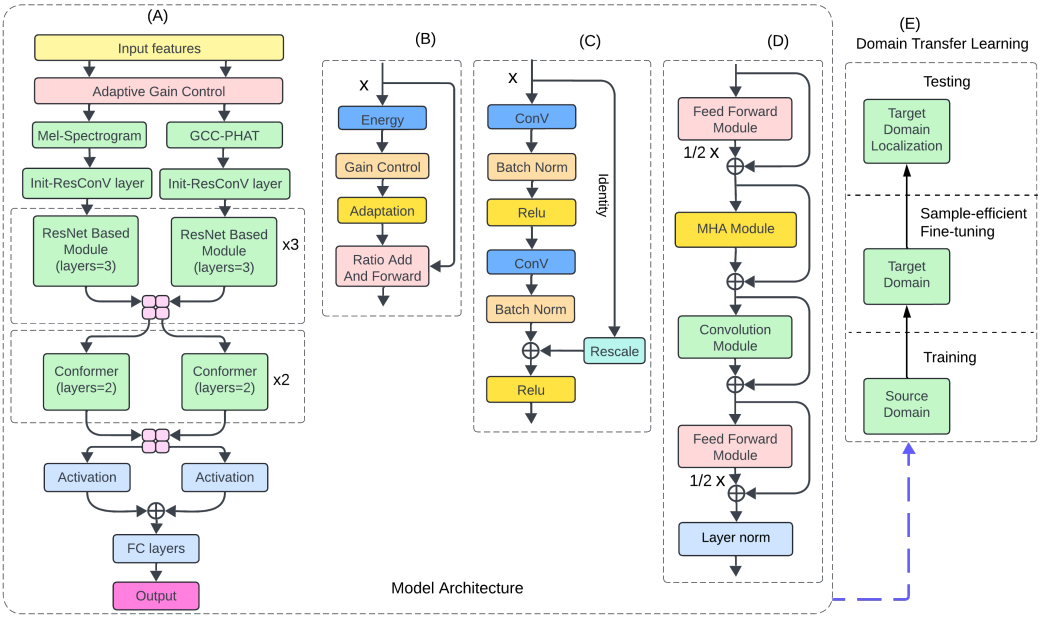

Fig. 1: (A) The Proposed ACA-Net Architecture. (B) The Adaptive Gain Control Layer. (C) The Rescaled ResNet Based Module. (D) The Conformer Block. (E) The Domain Transfer Learning Framework.

to a target domain with high performance using only a few labeled samples. This reduces the reliance on extensive labels, addressing the data scarcity issues in underwater acoustics.

Our proposed network achieves superior performance on the SWellEx-96 dataset, outperforming current SOTA solutions without the need for any additional data. The source-to-receiver distance (range) estimation task is formulated as a regression problem, which presents a greater challenge compared to the more commonly used classification approach [15], [25]. The results highlight the effectiveness of our method, particularly in handling complex multi-channel environments, noisy underwater acoustic conditions, and domain adaptation.

II. Proposed Methodology

A. Localization Model

The proposed ACA-Net is designed based on the improved event-independent network multi-branch architecture [26], [27]. In our work, we enhance the model by integrating ResNet-based blocks with rescaled residual connections and adding Conformers [28], [29] to improve the capture of temporal dependencies. This implementation introduces a multi-scale attention mechanism to better capture local and global features across channels. To reduce feature space complexity and mitigate overfitting, we apply max-pooling and dropout after each block. The dual-branch model learns log-mel spectrogram features in one branch and GCC-PHAT features in the other. It also employs learnable shared weights in the middle (2×2) windows. This approach enhances independent feature learning in each branch while still facilitating the discovery of interdependencies. The final layer is a multilayer perceptron of fully connected layers that outputs a single neuron predicting

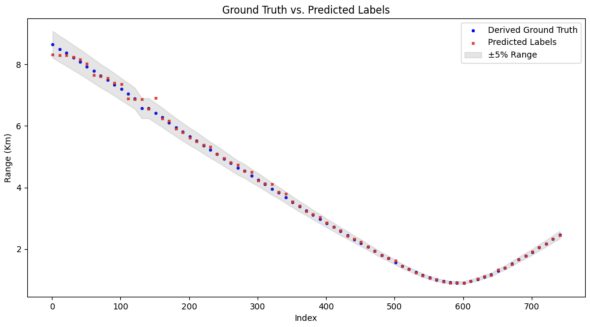

Fig. 2: In-domain S5 VLA. The x-axis "Index" represents test segment indices.††

location across the time domain, treating range estimation as a regression problem. The base network size is $40,502,545$ parameters and its architecture is illustrated in Figure 1.

B. Adaptive Gain Control (AGC) Layer

The AGC layer is introduced to dynamically adjust the signal amplitude during processing the input features. Implemented as a dynamic control layer, the AGC adaptively maintains consistent energy levels across input features, which enhances the robustness of subsequent processing stages. The AGC calculates the input's energy and uses it to compute a gain factor that adjusts the input to a target energy level. To ensure numerical stability, a small constant, $\epsilon = 10^{-6}$, is added to the denominator of the function. The gain ensures smooth

††For better visualization, every 10th indexed test segment is plotted.

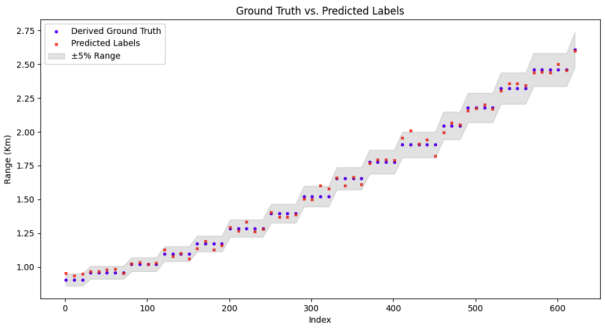

Fig. 3: Cross-domain Doppler Effect with 30% labels fine-tuning. The x-axis "Index" represents test segment indices.††

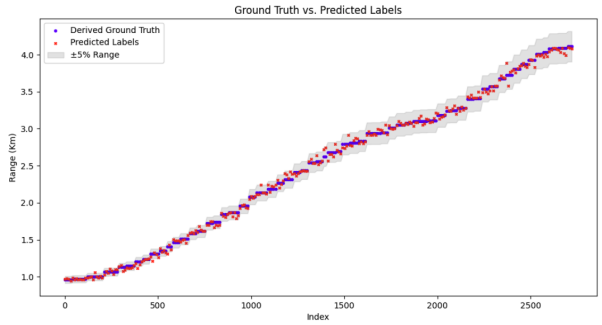

Fig. 4: Cross-domain S5 vs. S59 VLA with 30% labels fine-tuning. The x-axis "Index" represents test segment indices.††

transitions and prevents abrupt changes that could disrupt learning, without requiring input normalization or rescaling, unlike many other methods [19]–[21]. The energy per sample (E) and the AGC output ($x_{adaptive}$) are computed as:

$$E = \frac{1}{N}\sum_{i=1}^{N} x_i^2 \tag{1}$$

$$x_{adaptive} = \left[\left(\frac{E_{\text{target}}}{E+\epsilon} - 1\right)\alpha + 1\right] x \tag{2}$$

where N is the length of the input x, E_{target} is the defined target energy, and α is the adaptation rate. We conducted a grid search for the best parameters: E_{target} = 1.0 and α = 0.2.

All experiments are trained using supervised learning with backpropagation and mean squared error (MSE) loss, optimized with Adam (batch size 32, learning rate 10^{-4}), on an NVIDIA RTX 3090 GPU. The MSE is defined as:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{3}$$

where N is the total segments, and y_i and $\hat{y}_i$ are the ground truth and prediction at each segment index respectively.

III. Data

This study uses data from the S5 and S59 events of the SWellEx-96 experiment, which include Vertical Linear Array (VLA), Tilted Linear Array (TLA), Horizontal Linear Array (HLA) North, and HLA South. See [16] for more details.

Audio signals are segmented into non-overlapping one-second clips for all datasets. Each segment is labeled based on the closest "Duration" in minutes, as the original ground-truth range measurements are recorded once per minute. Log-mel spectrogram and GCC-PHAT features are extracted following the code and methods from [27], [30]. We use a Short Time Fourier Transform (STFT) with a "Hann" window of length 40 samples and 50% overlap to process spectrograms while capturing phase-difference cross-correlations between channel pairs to produce GCC-PHAT features.

TABLE I: In-domain S5 VLA – ACA-Net ablation study

Mel-spectro-gram feature	ResNet Module	GCC-PHAT feature	Con-former	AGC Layer	MAE (km)	$P_{CL-5\%}$ (%)
✓	✓		✓	✓	0.131	77.31
✓	✓	✓		✓	0.129	82.53
✓	✓	✓	✓		0.064	96.93
✓	✓	✓	✓	✓	**0.045**	**99.20**

TABLE II: In-domain S5 VLA - Range estimation comparison

	MAE (km)	MSE (km)	$P_{CL-5\%}$ (%)
MFP [17]	1.73	–	–
CNN-r [17]	1.40	–	–
CPA-DDA-UNET [22]	0.5976	–	–
FEAST [18]	0.5277	–	–
Encoder-MLP [19]	–	0.22	–
Siamese-SSL [20]	–	0.1207	–
Time-Freq-CPC [21]	–	0.11	–
MLF-TransCNN [23]	0.2718	–	–
RA-CAE-SSL [25]	0.0584	–	65.50
ACA-Net (Ours)	**0.0452**	**0.00546**	**99.20**

IV. Performance Metrics

In our experiments, we establish performance benchmarks with mean absolute error (MAE) and Probability of Credible Localization (P_{CL}). The P_{CL} metric measures the accuracy by considering predictions within a specified error limit. We evaluate the localization at a 5% error. They are defined as:

$$P_{CL-5\%} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(i) \tag{4}$$

where:

$$\mathbb{I}(i) = \begin{cases} 1, & \text{if } \frac{|y_i - \hat{y}_i|}{y_i} \times 100\% \leq 5\% \\ 0, & \text{otherwise} \end{cases}, \tag{5}$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{6}$$

where N is the total segments, y_i is the actual distance, and $\hat{y}_i$ is the predicted distance for the i-th index segment.

Additionally, Table II reports the MSE (Eq. 3) for comparison with methods that only provide benchmarks for MSE.

V. Experimental Setup

Our experiments include four in-domain setups for the VLA, TLA, HLA North, and HLA South arrays in S5 events, an ablation study, and two cross-domain setups (S5 VLA incoming vs. receding sound and S5 VLA vs. S59 VLA). In cross-domain experiments, we evaluate the network using zero-shot adaptation and fine-tuning with just 15% and 30% of target-domain data for sample-efficient domain transfer.

A. In-Domain Experimental Setups

For the in-domain S5 event VLA, TLA, HLA North, and HLA South, the data is evenly split into 6 folds: 4 for training, 1 for validation, and 1 for testing, ensuring balanced sample distribution. The process of constructing the 6-fold dataset is:

$$(x_i, y_i) \quad \forall i : fold = \mod (i, 6) \tag{7}$$

where $fold$ is the fold number of the segment, x_i is the ith segment in the dataset, and y_i is the corresponding label.

The VLA and TLA datasets each contain 4,500 total segments, covering 75 minutes of recorded sound. Similarly, the HLA North and HLA South datasets each comprise 3,000 total segments, spanning 50 minutes.

B. Cross-Domain Experimental Setups

In both cross-domain setups, we selected 0%, 15%, and 30% random segments from the test domain in each fine-tuning experiments respectively and include the results in Table IV.

1) Doppler Effect - S5 event VLA Approaching versus Receding Sound (Doppler): Approaching sound data from 0 to 59 minutes is used for training as the source domain, while receding sound data from 60 to 75 minutes (900 segments) serves as the test set in the target domain. This setup emulates real-world Doppler effects and assesses the model's generalization capabilities. The maximum Doppler shift is given by:

$$\Delta f = \frac{\pm 2.51}{1500} f_i \approx \pm 1.67 \times 10^{-3} f_i \tag{8}$$

where f_i is the source frequency. This corresponds to an approximate shift in the pilot tones of ± 0.082 to ± 0.67 Hz.

2) S5 VLA versus S59 VLA (S5 vs. S59): The model is trained on the S5 event VLA dataset and tested on the S59 event VLA dataset. Event S59 also involves a source towed along an isobath but includes an interferer nearby. The S59 event includes 65 minutes of audio, totaling 3900 one-second segments. This is useful for examining how a loud interferer affects the localization of a quiet target, presenting a greater challenge to the model in the target domain.

VI. Results

We conduct an ablation study to evaluate the effectiveness of key layer components and the impact of different selective feature sets as shown in Table I. It allows us to learn that the Conformer layer is the greatest contribution to our model, followed by GCC-PHAT feature and the AGC layer.

Table II compares our model with recent methods, showing that ACA-Net achieves SOTA performance on the SwellEx-96 S5 event VLA, as illustrated in Figure 2. Notably, the RA-CAE-SSL method [25] applies selective band-pass filters based on known frequencies to reduce noise and frames range estimation as a 100-class classification task within a narrow test range (0.903–2.608 km). Although it outperforms other SOTA methods, its dependence on prior frequency knowledge and fixed range bins limits scalability to unknown environments and makes it unsuitable for the regression task in this study. Compared to the more relevent current SOTA regression method MLF-TransCNN [23], our network achieves a significantly lower MAE of 0.0452 km versus 0.2718 km.

TABLE III: ACA-Net - In-domain setup performances

	MAE (km)	$P_{CL-5\%}(\%)$
S5 event VLA	0.045	99.20
S5 event TLA	0.044	98.67
S5 event HLA North	0.024	84.20
S5 event HLA South	0.037	81.80

(∗) All of our established benchmarks are SOTA performances.

TABLE IV: ACA-Net - Cross-domain setup performances

Setup	**Data Used (%)**	**MAE (km)**	$P_{CL-5\%}(\%)$
Doppler	Zero-shot 0% (§)	0.078	59.11
	Fine-tuned 15% (§)	0.032	92.94
	Fine-tuned 30% (§)	**0.025**	**98.57**
	Trained-scratch 15% (†)	0.113	44.44
	Trained-scratch 30% (†)	0.057	77.46
S5 vs. S59	Zero-shot 0% (§)	0.441	15.21
	Fine-tuned 15% (§)	0.070	81.06
	Fine-tuned 30% (§)	**0.032**	**96.92**
	Trained-scratch 15% (†)	0.128	54.54
	Trained-scratch 30% (†)	0.061	89.27

(§) Source domain pre-trained; fine-tuned with test domain data (%).
(†) No pre-training; trained with test domain data (%) from scratch.

Similar trends, such as lower performance on long-range estimation, are observed for the TLA and HLAs, but figures are omitted due to the page limit. Table III displays results from our various in-domain experiments. We are among the first to conduct experiments on the S5 TLA, HLA North and South. ACA-Net demonstrates robust performance across various multi-channel array configurations. The HLAs arrays exhibit lower $P_{CL-5\%}$ scores, indicating reduced accuracy due to fewer data samples (50 minutes) compared to the VLA and TLA arrays (75 minutes). Despite this, they achieve slightly better MAE, likely because their recordings cover narrower range intervals in those 50 minutes (See [16] for more details). Meanwhile, the TLA and VLA, with equal durations and identical range coverage, show consistently comparable performance as reported in Table III.

Additionally, the results in Table IV highlight the effectiveness of source-domain pre-training for cross-domain adaptation. As test-domain data decreases, the benefits of pre-training become more pronounced, especially in the Doppler Effect experiment where test data is more limited. The source-pretrained network achieves zero-shot performance only slightly below that of training from scratch with 30% of the target domain data. Similarly, in the S5 vs. S59 setting,

fine-tuning the source-pretrained model with only 15% of test domain data achieves performance on par with training from scratch with 30% of the data. Table IV and Figures 3 and 4 show that our model is robust to the Doppler Effect and loud acoustic interference, suggesting generalization capability in domain adaptation with minimal training or fine-tuning.

VII. Conclusions

Our proposed ACA-Net outperforms SOTA methods on the S5 event VLA, TLA, HLA North, and HLA South dataset, achieving strong cross-domain adaptation with minimal fine-tuning in the S5 VLA vs. S59 VLA and Doppler Effect experiments, suggesting a strong generalization. It shows the benefits of using multiple features via a multi-branch architecture and the ability to adapt to a new array dataset efficiently. While the network achieves superior performance with a range of arrays and events, it has not yet been evaluated under a wider variety of oceanic conditions. Future research will focus on creating a dynamic feature extraction layer to process signals from various arrays, allowing localization independent of array structure, and will involve testing the network on additional underwater datasets. Additional acoustic information may also be used to enhance training and reduce data requirements through a more physics-informed ML approach.

Acknowledgment

The work on this paper was supported by the Office of Naval Research (Grant No. N00014-21-1-2790).

References

[1] M. Dong, H. Li, Y. Qin, Y. Hu, and H. Huang, "A secure and accurate localization algorithm for mobile nodes in underwater acoustic network," *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108157, 2024.

[2] U. T. Jurado, P. Wilson, P. Blondel, A. Bartin, and G. Walker-Doyle, "Underwater localization system for marine seismic airgun arrays validated through robotics," *International Journal of Intelligent Robotics and Applications*, pp. 1–13, 2025.

[3] A. Baggeroer, W. Kuperman, and P. Mikhalevsky, "An overview of matched field methods in ocean acoustics," *IEEE Journal of Oceanic Engineering*, vol. 18, no. 4, pp. 401–424, 1993.

[4] M. B. Porter, "The kraken normal mode program," *Naval Research Lab Washington DC, Report No. NRL/MR/5120—92-6920*, 1992.

[5] N. Kasthuri, S. Balambigai, and S. Yuvashree, "Source localization for underwater acoustics using esprit algorithm," *IOP Conference Series: Materials Science and Engineering*, vol. 1055, 2021.

[6] R. Wang, M. Ding, L. Yue, and H. Liu, "Design of underwater acoustic localization device based on beamforming," *2024 8th International Conference on Electrical, Mechanical and Computer Engineering (ICEMCE)*, pp. 1055–1058, 2024.

[7] M. Rezzouki, G. Ferré, G. Terrasson, and A. Llaria, "Net fishing localization: Performance of tdoa-based positioning technique in underwater acoustic channels using chirp signals," *2024 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, 2024.

[8] T. Guo, Y. Song, Z. Kong, E. Lim, M. López-Benítez, F. Ma, and L. Yu, "Underwater target detection and localization with feature map and cnn-based classification," *2022 4th International Conference on Advances in Computer Technology, Information Science and Communications (CTISC)*, pp. 1–8, 2022.

[9] Y. Z. Nga, Z. Rymansaib, A. A. Treloar, and A. J. Hunter, "Automated recognition of submerged body-like objects in sonar images using convolutional neural networks," *Remote Sensing*, 2024.

[10] Y. Xu, P. Liu, L. Wang, and J. Ma, "Main body shape optimization of non-body-of-revolution underwater vehicles by using cnn and genetic algorithm," *Ocean Engineering*, 2024.

[11] N. Davari and A. P. Aguiar, "Real-time outlier detection applied to a doppler velocity log sensor based on hybrid autoencoder and recurrent neural network," *IEEE Journal of Oceanic Engineering*, vol. 46, pp. 1288–1301, 2021.

[12] W. Zhang, X. tong Yang, C. Leng, J. Wang, and S. Mao, "Modulation recognition of underwater acoustic signals using deep hybrid neural networks," *IEEE Transactions on Wireless Communications*, vol. PP, pp. 1–1, 2022.

[13] S. Feng and X. Zhu, "A transformer-based deep learning network for underwater acoustic target recognition," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[14] Q. T. Vo and D. K. Han, "Underwater acoustic signal classification using hierarchical audio transformer with noisy input," in *2023 IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP)*, 2023, pp. 1–6.

[15] H. Niu, X. Li, Y. Zhang, and J. Xu, "Advances and applications of machine learning in underwater acoustics," *Intelligent Marine Technology and Systems*, vol. 1, no. 1, p. 8, 2023.

[16] J. Murray and D. Ensberg, "The swellex-96 experiment," 1996. [Online]. Available: https://swellex96.ucsd.edu/index.htm

[17] R. Chen and H. Schmidt, "Model-based convolutional neural network approach to underwater source-range estimation," *The Journal of the Acoustical Society of America*, vol. 149, no. 1, pp. 405–420, 01 2021.

[18] J. Chi, X. Li, H. Wang, D. Gao, and P. Gerstoft, "Sound source ranging using a feed-forward neural network trained with fitting-based early stopping," *The Journal of the Acoustical Society of America*, vol. 146, no. 3, pp. EL258–EL264, 09 2019.

[19] X. Zhu, H. Dong, P. Salvo Rossi, and M. Landrø, "Feature selection based on principal component regression for underwater source localization by deep learning," *Remote Sensing*, vol. 13, no. 8, 2021.

[20] H. Wen, C. Yang, D. Dou, L. Xu, and Y. Jiao, "Underwater source ranging by siamese network aided semi-supervised learning," *JASA Express Letters*, vol. 3, no. 9, p. 094803, 09 2023.

[21] X. Zhu, H. Dong, P. S. Rossi, and M. Landrø, "Time-frequency fused underwater acoustic source localization based on contrastive predictive coding," *IEEE Sensors Journal*, vol. 22, no. 13, pp. 13 299–13 308, 2022.

[22] X. Wang, H. Zhou, J. He, B. Zhang, and R. Tang, "A multi-scale coordinate embedded dual-path u-net for shallow-sea source localization," in *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering (CAICE)*. Association for Computing Machinery, 2024, p. 67–73.

[23] J. He, B. Zhang, P. Liu, X. Li, L. Wang, and R. Tang, "Effective underwater acoustic target passive localization of using a multi-task learning model with attention mechanism: Analysis and comparison under real sea trial datasets," *Applied Ocean Research*, vol. 150, p. 104072, 2024.

[24] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[25] P. Jin, B. Wang, L. Li, P. Chao, and F. Xie, "Semi-supervised underwater acoustic source localization based on residual convolutional autoencoder," *EURASIP Journal on Advances in Signal Processing*, vol. 2022, no. 1, p. 107, 2022.

[26] Y. Cao, T. Iqbal, Q. Kong, F. An, W. Wang, and M. D. Plumbley, "An improved event-independent network for polyphonic sound event localization and detection," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 885–889.

[27] Q. T. Vo and D. K. Han, "Resnet-conformer network with shared weights and attention mechanism for sound event localization, detection, and distance estimation," DCASE2024 Challenge, Tech. Rep., 2024. [Online]. Available: https://dcase.community/documents/challenge2024/technical_reports/DCASE2024_Vo_63_t3.pdf

[28] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proceedings of Interspeech*, 2020, pp. 5036–5040.

[29] Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, and Q. Ye, "Conformer: Local features coupling global representations for visual recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 367–376.

[30] D. A. Krause, A. Politis, and A. Mesaros, "Sound event detection and localization with distance estimation," in *2024 32nd European Signal Processing Conference (EUSIPCO)*, 2024, pp. 286–290.