One2ThreeNet: An Automatic Microscale-Based Modulation Recognition Method for Underwater Acoustic Communication Systems

Jingjing Wang, *Member, IEEE*, Zihao Huang, Wei Shi, *Member, IEEE*, and Shiwen Mao, *Fellow, IEEE*

***Abstract*— Automatic modulation recognition (AMR) technology enables receivers to automatically recognize the modulation type of the received signal for correct demodulation of the received data, but there are still many shortcomings to be addressed. To achieve accurate and efficient AMR, this paper proposes a data augmentation method for AMR, which can increase the amount of data by seven times and solve the problem of a small sample size more effectively than the existing methods. In addition, this paper proposes a concept of microscale, rationalizes the underwater acoustic signal into time series, and proposes a temporal feature extractor named One2Three block, which can extract temporal features of signals from three microscales. Finally, a spatial feature extractor named the Dual-Stream squeeze-and-excitation (SE) block is designed to abstract and synthesize more advanced spatial features for AMR. The recognition accuracy of the proposed method is verified with eight commonly used modulation modes in underwater acoustic communications on the datasets collected in the South China Sea and the Yellow Sea. The results show that the proposed method can achieve a recognition accuracy of 99% with a lower time and space complexity, and has high robustness to noisy data.**

***Index Terms*— Automatic modulation recognition, convolutional neural network, data augmentation, deep learning, One2ThreeNet.**

I. Introduction

IN AN underwater acoustic communication system, the transmitting end usually uses the adaptive modulation coding technology to transmit data according to the current channel condition [1]. However, the complex ocean noise can severely interfere with the handshake signal of both sides of the communication link, causing the receiver to fail to confirm the modulation mode of the received signal, which can result in communication failure [2], [3]. As an important step in blind signal processing, automatic modulation recognition (AMR) enables receivers to automatically identify the received signal's modulation type for accurate signal demodulation [4]. Therefore, AMR has high significance for the research on cognitive and software-defined radio, and in particular, for underwater communication systems [5], [6].

With the continuous enhancement of computing power, deep learning (DL) has achieved remarkable success in many fields, such as image recognition [7], object detection [8], and sentiment analysis [9]. By fitting a complex and powerful function with many parameters, deep neural networks can automatically learn how to combine low-level features to obtain abstract high-level features, and then accomplish complex tasks that are difficult to perform by traditional machine learning-based methods [10]. Due to the strong learning and fitting abilities of DL, it has been gradually applied to AMR in recent years [11], [12]. In the field of AMR, commonly-used deep learning models mainly include recurrent neural networks (RNNs) and convolutional neural networks (CNNs) [13]. The RNNs have largely been used to process time-series data, for the purpose of extracting temporal features by learning the correlation between the input sequences in the temporal dimension. In contrast, CNNs have been mostly used in the field of image processing, and their core component is a convolution operator, which enables a CNN network to extract spatial features of the input image by fusing the spatial and channel information in the local receptive field of each layer [14].

In the early stage, researchers focused on the application of CNNs to AMR. For example, the authors in [15] developed a four-layer CNN for AMR. However, due to the lack of samples, their model could not achieve the ideal recognition performance. In [16], an AMR method based on the residual network (ResNet) was proposed. Under the signal-to-noise ratio (SNR) of 6 dB, this method could achieve a recognition accuracy of 90% for binary phase shift keying (2PSK), quaternary phase shift keying (4PSK), 16-ary amplitude phase shift keying (16APSK), and 32-ary amplitude phase shift keying (32APSK). However, CNNs have a poor learning ability for temporal features. When signal sampling points are simply reconstructed into a feature map and fed to the CNN input, it will be challenging to achieve a good recognition

Manuscript received 15 February 2023; revised 10 June 2023 and 17 December 2023; accepted 14 February 2024. Date of publication 12 March 2024; date of current version 14 August 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62171246, Grant U1806201, and Grant 62101298; and in part by the Major Basic Research Project of Natural Science Foundation of Shandong Province under Grant ZR2021ZD12 and Grant ZR2021MF130. The associate editor coordinating the review of this article and approving it for publication was Y. Shu. *(Corresponding author: Zihao Huang.)*

Jingjing Wang and Wei Shi are with the College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao, Shandong 266061, China (e-mail: wangjingjing@qust.edu.cn; shiwei6670595@126.com).

Zihao Huang is with the Pen-Tung Sah Institute of Micro-Nano Science and Technology, Xiamen University, Xiamen, Fujian 361000, China (e-mail: Zihao_yolo@qq.com).

Shiwen Mao is with the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849 USA (e-mail: smao@ieee.org).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2024.3371226.

Digital Object Identifier 10.1109/TWC.2024.3371226

performance. To address these challenges, recent studies have used a method of converting signals into images based on expert experience and then employed the CNN models for AMR. In [17], an AMR method for converting signals to eye diagrams and training a CNN was proposed. However, this method can be applied only to the recognition of baseband signals and thus has a limited application value. In [18], a CNN-based AMR model was developed to learn spatial features of the cyclic spectrum, and sparse filtering criteria were defined to enhance the recognition performance. The results indicated that 2PSK, 4PSK, 2FSK, and 4FSK signals can be successfully recognized.

The aforementioned studies have transformed the AMR problem into an image recognition problem, but most of them used specific transformation schemes for specific problems and required complex preprocessing of data. Therefore, their proposed solutions can be regarded as another form of artificial feature extraction, which will inevitably result in the loss of many crucial temporal features. To overcome this problem, recent studies have combined CNNs and RNNs to extract both spatial and temporal features to obtain a more comprehensive feature set. In [19], a double-channel structure, consisting of a CNN and a long-short term memory (LSTM) network to estimate modulation type. The recognition accuracy of this method was approximately 90% at a high SNR. A three-channel network structure for AMR named multi-channel convolutional long short-term deep neural network (MCLDNN) was proposed in [20], which can effectively extract spatial and temporal features and achieve a recognition accuracy of 92%. Due to the strong temporal features in underwater acoustic signals, it would be illogical to extract spatial features before temporal features. Therefore, this paper proposes an AMR method called recurrent and convolutional neural network (R&CNN) [21], which extracts temporal features first and then spatial features and achieved a recognition accuracy of more than 99%. However, the proposed network has high time and space complexity and lacks interpretability and generality.

Deep Learning's performance largely hinges upon the availability of sufficient high-quality labeled data. However, the datasets obtained in actual communications scenarios usually contain only a small number of samples. Therefore, it is particularly important to perform data augmentation on the dataset. There are a few studies on data enhancement in the AMR domain. In [22], the author proposed a conditional variational auto-encoder (CVAE)-enhanced learning model for AMR, since the CVAE-generated data maintains more key features which help improve the classification accuracy, via the feedback link from the classifier to the CVAE decoder, the dataset enhanced by this method has a higher classification accuracy than the data set enhanced by other traditional CNN algorithms. In [23], the author proposed a data augmentation scheme with the conditional generative adversarial network (CGAN) for AMR, which can synthesize high-quality, labeled modulation data from a small set of available real data. However, the above methods are all targeted at baseband signals, and to the best of our knowledge, there is no effective data enhancement method for frequency-band signals at present.

In this paper, we address the AMR problem of underwater acoustic communication, with a DL-based method. The main contributions of this study are as follows:

1. To solve the performance limitation problem of DL-based algorithms under scarce samples, we propose an effective data augmentation method, which can be used in automatic modulation recognition of underwater acoustic communication, and the size of a dataset can be increased by seven times;

2. To rationalize the received signal into a time series, we proposed the concept of microscale of an underwater acoustic signal, which is instrumental for the subsequent extraction of temporal signal features;

3. A temporal feature extractor named One2Three is proposed. The One2Three block converts temporal features of a one-dimensional signal into three-channel feature maps from three microscales to extract temporal features while facilitating the spatial feature extraction by using a CNN;

4. A spatial feature extractor named Dual-Stream squeeze-and-excitation (SE) is proposed to construct a dual-channel attention mechanism network, which can not only effectively broaden the network width, but also improve the feature map quality. Combined with the One2Three block, this module can effectively extract the temporal and spatial features of underwater acoustic signals, thus achieving efficient and reliable AMR;

5. The proposed method is verified on the datasets collected from the South China Sea and the Yellow Sea. The experimental results show that the proposed method can accurately identify eight modulation modes, including DSSS, 2FSK, 4FSK, 2PSK, 4PSK, 16QAM, 64QAM, and OFDM. The recognition accuracy of the proposed method can reach 99%. In addition, the parameters of this model are only 0.28 M and floating-point operations (FLOPs) are only 7.02 M, indicating its extremely low space complexity and time complexity.

The remainder of the paper is organized as follows. Section II presents the mathematical model and signal preprocessing method. Section III introduces the One2ThreeNet model. Section IV describes the data augmentation method in detail. Section V explains the proposed model structure and introduces the datasets. Section VI verifies the effectiveness of the proposed method through a variety of experiments. Finally, Section VII concludes the paper.

II. Underwater Acoustic Communication System Model

To measure an underwater acoustic signal, we designed an underwater acoustic communication system and conducted extensive underwater acoustic communication experiments in the South China Sea and the Yellow Sea of China. The underwater acoustic communication equipment used in the experiments is shown in Fig. 1.

In the experiments, the receiving and transmitting sides were on the ports of two test ships. The transmitter first converted a digital signal to an analog signal using a digital-to-analog converter (DAC) and then amplified the converted signal by a power amplifier (PA), and finally, converted the amplified electrical signal to an acoustic signal through a transducer. The receiver also used transducers to convert the received

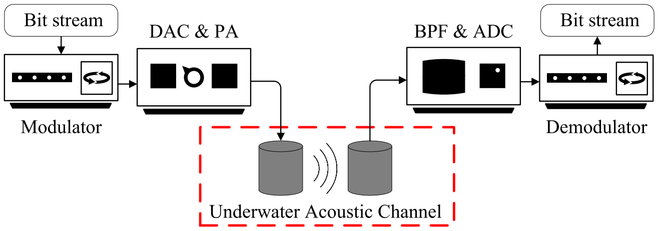


Fig. 1. The block diagram of the underwater acoustic communication equipment used in the experiments.

acoustic signal into an electrical signal and employed a band-pass filter (BPF) to filter the obtained electrical signal. Finally, the sampled underwater acoustic signal was obtained by an analog-to-digital converter (ADC).

An underwater acoustic communications system can be mathematically described as follows:

$$y(t) = Ae^{j(2\pi f_0 t+\theta_0)} \int_{-\infty}^{+\infty} s(\tau)g(\tau - t)h(t-\tau+\zeta)d\tau + a(t) + I(t), \quad (1)$$

where $s(t)$ and $y(t)$ denote the transmitted and received signals, respectively; A denotes the amplitude gain of the received signal; f_0 and θ_0 are the frequency offset and phase shift at the receiving end, respectively; $g(t)$ represents the function of pulse shaping filter; $h(t)$ is the impulse response of an underwater acoustic channel; ζ denotes the time synchronization error at the receiving end; $a(t)$ and $I(t)$ are the additive white Gaussian noise and the additive impulse noise, respectively.

This study assumes that the receiver has completed frame synchronization and that its sampling rate is f_s. After sampling an analog signal $y(t)$ with a time interval of $1/f_s$, a discrete signal $y[n]$ with a length of N is obtained. The main objective of AMR can be expressed as accurately recognizing the modulation type when only a discrete signal $y[n]$ is received.

III. ONE2THREENET

A. Signal Microscale

In digital communications, changes in the carrier frequency, phase, and amplitude of a signal correspond to the information carried by symbols. Typically, the types of symbols used in a certain modulation mode are fixed. From a macro perspective, a signal represents a collection of random, unrelated symbols, lacking the characteristics of time series. It should be noted that the signals of different modulation types have different phase and amplitude changes, which can be captured on the scale of a single symbol. So, from a micro perspective, each sampling point constituting a single symbol can be regarded as a time series with certain autocorrelation. This study considers three microscales, which are as follows:

1) Double-symbol scale P_1: Features of modulation types are reflected in modifications of adjacent symbols (e.g., MPSK signals and MFSK signals). In MPSK signals, the phase between adjacent symbols changes. In MFSK signals, the frequency between adjacent symbols may also change;

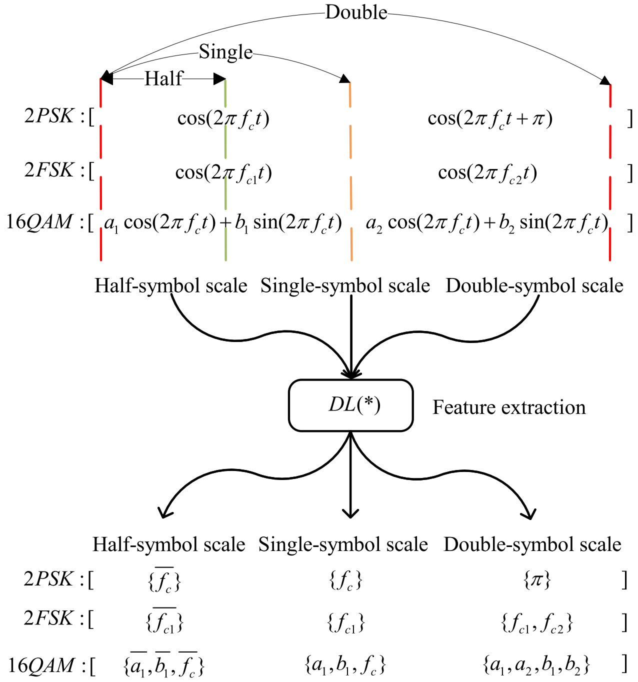

Double
Single
Half
$2PSK$: [$\cos(2\pi f_c t)$ $\cos(2\pi f_c t+\pi)$]
$2FSK$: [$\cos(2\pi f_{c1} t)$ $\cos(2\pi f_{c2} t)$]
$16QAM$: [$a_1\cos(2\pi f_c t)+b_1\sin(2\pi f_c t)$ $a_2\cos(2\pi f_c t)+b_2\sin(2\pi f_c t)$]
Half-symbol scale Single-symbol scale Double-symbol scale
$DL(*)$
Feature extraction
Half-symbol scale Single-symbol scale Double-symbol scale
$2PSK$: [$\{\overline{f_c}\}$ $\{f_c\}$ $\{\pi\}$]
$2FSK$: [$\{\overline{f_{c1}}\}$ $\{f_{c1}\}$ $\{f_{c1}, f_{c2}\}$]
$16QAM$: [$\{\overline{a_1}, \overline{b_1}, \overline{f_c}\}$ $\{a_1, b_1, f_c\}$ $\{a_1, a_2, b_1, b_2\}$]

Fig. 2. Feature extraction at different microscales.

2) Single-symbol scale P_2: At this microscale, symbols with different modulation modes have different waveforms in the time domain. In addition, the envelope of time-domain waveforms of a single symbol (e.g., MQAM signals) reflects the modulation characteristics of signals employing amplitude modulation. The value of P_2 is determined by:

$$P_2 = f_s / R_s, \quad (2)$$

where f_s is the sampling frequency of a receiver, and R_s is the symbol rate of the received signal.

3) Half-symbol scale P_3: A raised cosine filter has often been used to achieve pulse shaping in underwater acoustic communication systems. The raised cosine function is an even function, fine-grained temporal features of a signal can be extracted from the half-symbol scale.

Due to the temporal features of the signal corresponding to the microscale, we draw on the idea of Short-time Fourier Transform (STFT), use a rectangular window to segment the signal into discrete symbols and give the time domain expression of each symbol, as shown in Fig. 2.

As shown in Fig. 2, signals at different microscales are input into the neural network to obtain features at different microscales. To facilitate understanding, we choose 2PSK, 2FSK, and 16QAM signals for example. 2PSK, 2FSK, and 16QAM signal expressions are shown below respectively, and each expression contains two symbol lengths.

$$\begin{aligned} 2PSK &: [\cos(2\pi f_c t), \cos(2\pi f_c t+\pi)], \\ 2FSK &: [\cos(2\pi f_{c1} t), \cos(2\pi f_{c2} t)], \\ 16QAM &: [a_1 \cos(2\pi f_c t) + b_1 \sin(2\pi f_c t), \\ &\quad a_2 \cos(2\pi f_c t) + b_2 \sin(2\pi f_c t)]. \end{aligned} \quad (3)$$

In addition, we abstract the deep neural network as a feature extraction function DL $(*)$ to intuitively explain which features

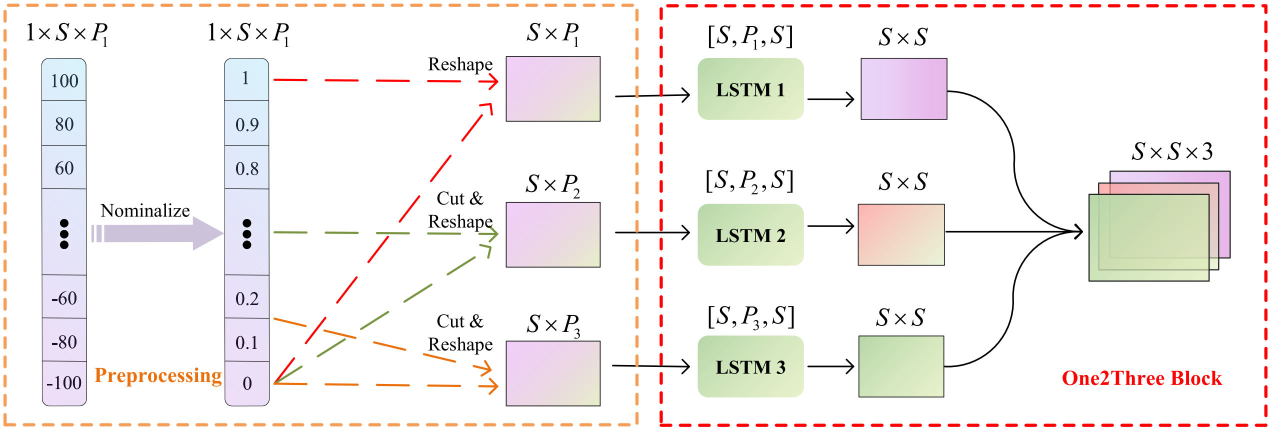


Fig. 3. Signal preprocessing and the network structure of the One2Three block.

can be extracted for recognition at different microscales. Taking the 2PSK signal as an example, it can be intuitively seen from the signal expression that the phase change π can be extracted under the double-symbol scale, the carrier f_c can be extracted under the single-symbol scale. Since the baseband signal shaping filter is an even function and the microscale of the half-symbol is smaller than that of the single-symbol, fine-grained features $\overline{f_c}$ related to f_c can be extracted under the half-symbol scale.

B. One2Three Block Structure

The long short-term memory (LSTM) model proposed by Hochreiter [24] represents an RNN variant. By constructing a long-term memory storage unit, this model mitigates the problems of gradient disappearance and gradient explosion caused by extremely long time series and enables the network to learn both short-term association features and long-term dependence relations [25]. Assume the input sequence corresponding to the t-th time step is denoted by $\mathbf{X}_t$; then the output vector $\mathbf{y}_t$ corresponding to this time step is as follows:

$$\begin{aligned}
\mathbf{Z}_f &= \sigma\left(\mathbf{W}_f \odot [\mathbf{h}_{t-1}, \mathbf{X}_t]\right), \\
\mathbf{Z}_i &= \sigma\left(\mathbf{W}_i \odot [\mathbf{h}_{t-1}, \mathbf{X}_t]\right), \\
\mathbf{Z}_o &= \sigma\left(\mathbf{W}_o \odot [\mathbf{h}_{t-1}, \mathbf{X}_t]\right), \\
\mathbf{Z} &= \tanh\left(\mathbf{W} \odot [\mathbf{h}_{t-1}, \mathbf{X}_t]\right), \\
\mathbf{C}_t &= \mathbf{Z}_f \odot \mathbf{C}_{t-1} + \mathbf{Z}_i \odot \mathbf{Z}, \\
\mathbf{h}_t &= \mathbf{Z}_o \odot \tanh\left(\mathbf{C}_t\right), \\
\mathbf{y}_t &= \sigma\left(\mathbf{W}' \odot \mathbf{h}_t\right), \qquad (4)
\end{aligned}$$

where $\mathbf{h}_{t-1}$ is the hidden layer state vector of the previous time step; the vector after $\mathbf{h}_{t-1}$ and $\mathbf{X}_t$ are concatenated constituting the input of the t-th time step; $\mathbf{W}_f$,$\mathbf{W}_i$,$\mathbf{W}_o$, and $\mathbf{W}'$ are the weight vectors, affect the forgetting, input, and output parts of the network, respectively; σ represents the sigmoid activation function; operator $\odot$ represents the dot multiplication; $\mathbf{C}_{t-1}$ represents the long-term memory of the previous time step; $\mathbf{C}_t$ represents the long-term memory at the current time step; $\mathbf{Z}_f$, $\mathbf{Z}_i$, and $\mathbf{Z}_o$ represent the gating weights of the forgetting, input, and output gates, respectively, they are mainly used to update the long-term memory of the network; $\mathbf{Z}_f$ controls how much information in $\mathbf{C}_{t-1}$ needs to be forgotten, $\mathbf{Z}_i$ controls how much information in $\mathbf{Z}$ needs to be updated into $\mathbf{C}_t$; $\mathbf{h}_t$ is the hidden layer state at the current time step, calculated by $\mathbf{Z}_o$ and $\mathbf{C}_t$, $\mathbf{h}_t$ is weighted by $\mathbf{W}'$ to get the final output of the network.

Considering the powerful temporal feature extraction ability of the LSTM model, this paper constructs a temporal feature extractor named the One2Three block, which converts a one-dimensional sequence into a three-dimensional feature map, signal preprocessing and the network structure of the One2Three block is shown in Fig. 3.

As shown in Fig. 3, before input into the One2Three block, the signal needs to be preprocessed and reconstructed, first, the signal amplitude is normalized to the range of [0,1] to reduce the impact of large signal amplitude differences on the network convergence speed and recognition accuracy. To obtain signals that conform to the input size of the LSTM model, this paper converts the preprocessed signal $\tilde{y}$ into three two-dimensional matrices, with dimensions of $S \times P_1$, $S \times P_2$, and $S \times P_3$, respectively, where S represents the number of time steps of the LSTM model.

Then, the two-dimensional matrices are fed into three LSTM models to extract time-series features of the signal at the three microscales, and three feature maps with a size of $S \times S$ are obtained. The number of units in each LSTM is S, and the number of output layers size in each LSTM is S, too, the input layers size of LSTM1, LSTM2, and LSTM3 are P_1, P_2, and P_3, respectively. Finally, the obtained maps are merged in the third dimension to obtain the resulting feature map with a size of $S \times S \times 3$, thus completing the transformation from a one-dimensional sequence to three-channel feature maps. In this paper, since the sampling rate of the received signal f_s is 0.336 MHz and the symbol rate R_s is 560 baud, P_2 is equal to 600 according to equation (2), P_1 is equal to 1200, and P_3 is equal to 300. S is set to 16.

The RGB images can be regarded as image representations in three color domains. Similarly, the feature map extracted by the One2Three block can be interpreted as a signal representation at double-symbol scale, single-symbol scale, and half-symbol scale.

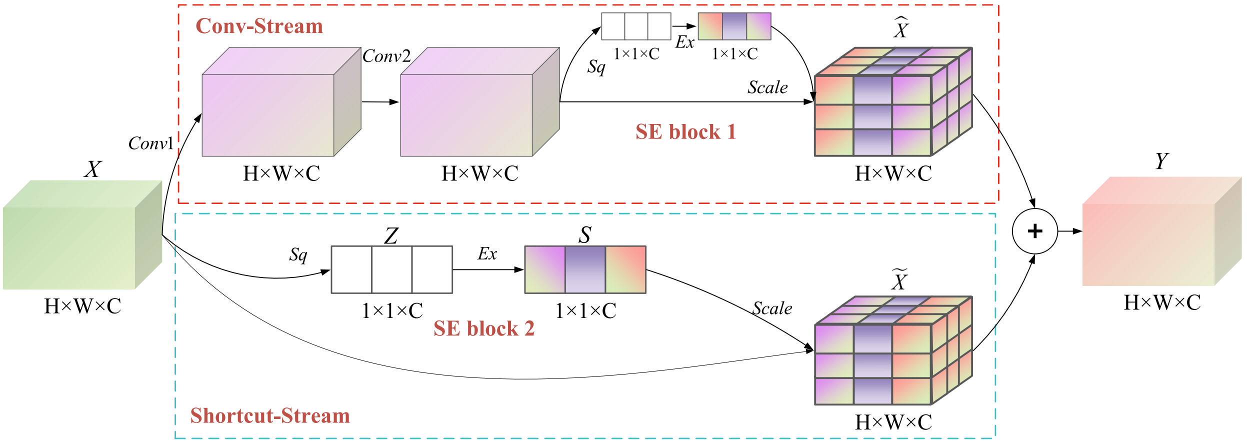


Fig. 4. The network structure of the proposed dual-stream SE block.

C. Dual-Stream SE Block Structure

After the received signal is transformed into a three-channel feature map by the One2Three block, a CNN model is used to extract the spatial features from the feature map and realize AMR.

The visual geometry group network (VGGNet) [26] and the inception model [27] show that increasing the network depth can significantly improve model performances, but an increased depth can also easily cause the gradient disappearance or gradient explosion problems. By using shortcut based on identity mapping, the deep residual network (ResNet) provides a shortcut for gradient propagation across different network layers, thus allowing to train deeper and more powerful networks [28]. In addition, the attention mechanism has achieved excellent results in many fields [29]; namely, it represents a higher level of abstraction and guides the model to learn more effective expressions. On this basis, the squeeze-and-excitation networks (SENet) [30] focuses on the performance comparison of channels in the feature map, and a squeeze-and-excitation block, called the SE block, is proposed. The SE block assigns the corresponding weight to each channel according to the channel importance to suppress invalid channels while improving the response of the effective channels.

In this study, we propose to add the SE block to the shortcut of the ResNet because this can not only expand the network width but also optimize the representation of the feature map of the ResNet. Therefore, this paper designs a dual-stream SE block based on the ResNet and SENet architecture, whose structure is shown in Fig. 4.

As shown in Fig. 4, in the Conv-Stream block, the input feature map is processed by two convolution layers to obtain a higher-level feature map. Then, the results are input into SE block 1 to obtain the feature map $\widehat{\mathbf{X}}$; the numbers of convolution kernels used in the two convolution layers are C and the size of convolution kernels are $H \times W$. The Shortcut-Stream block constructs a shortcut of gradient propagation, and the input feature map $\mathbf{X}$ is processed by the SE block 2 to obtain the feature map $\widetilde{\mathbf{X}}$. Finally, the output feature map $\mathbf{Y}$ is obtained by adding $\widehat{\mathbf{X}}$ and $\widetilde{\mathbf{X}}$. Both the Conv-Stream block and the Shortcut-Stream block use the SE blocks to optimize the response of different channels of the feature map.

In the Shortcut-Stream block, an SE block consists of Sq, Ex, and $Scale$ operations, which can be abstracted into three functions F_{sq}, F_{ex}, and F_{scale}, respectively, assuming that $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_C]$, $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$.The squeezed weight vector is $\mathbf{Z} \in \mathbb{R}^{C\times 1}$, and the c-th element in Z can be calculated by:

$$\mathbf{Z}_c = F_{sq}\left(\mathbf{x}_c\right) = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbf{x}_c(i,j). \tag{5}$$

The weight vector after excitation is $\mathbf{S} \in \mathbb{R}^{C\times 1}$, given by:

$$\mathbf{S} = F_{ex}\left(\mathbf{Z}, \mathbf{w}_1, \mathbf{w}_2\right) = \sigma\left(\mathbf{w}_2\delta\left(\mathbf{w}_1 \odot \mathbf{Z}\right)\right), \tag{6}$$

where σ represents the sigmoid activation function; operator $\odot$ represents the dot multiplication; δ is the Relu activation function; $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r}\times C}$, $\mathbf{W}_2 \in \mathbb{R}^{C\times\frac{C}{r}}$; r represents the compression rate of the SE block. The final output of the SE block $\widetilde{\mathbf{X}} = [\widetilde{\mathbf{x}}_1, \widetilde{\mathbf{x}}_2, \ldots, \widetilde{\mathbf{x}}_C]$, $\widetilde{\mathbf{X}} \in \mathbb{R}^{H\times W\times C}$ can be obtained by using $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_C]$ to weight $\mathbf{X}$, and the c-th element in $\widetilde{\mathbf{X}}$ can be calculated by:

$$\tilde{\mathbf{x}}_c = F_{scale}\left(\mathbf{x}_c, \mathbf{s}_c\right) = \mathbf{s}_c \odot \mathbf{x}_c. \tag{7}$$

The output feature map of the dual-stream SE block is calculated by:

$$\mathbf{Y} = \widehat{\mathbf{X}} + \widetilde{\mathbf{X}}, \tag{8}$$

where $\widehat{\mathbf{X}}$ and $\widetilde{\mathbf{X}}$ denote the feature maps of the Conv-Stream and Shortcut-Stream blocks, respectively.

To ensure consistency in feature map size before and after convolution, in the dual-stream SE block, the padding and sliding steps of the convolution operation are both equal to one.

IV. Data Augmentation

The performance of deep learning models relies on the availability of large amounts of high-quality data. However, due to the high acquisition cost of measured underwater acoustic signals, most of the existing deep learning-based

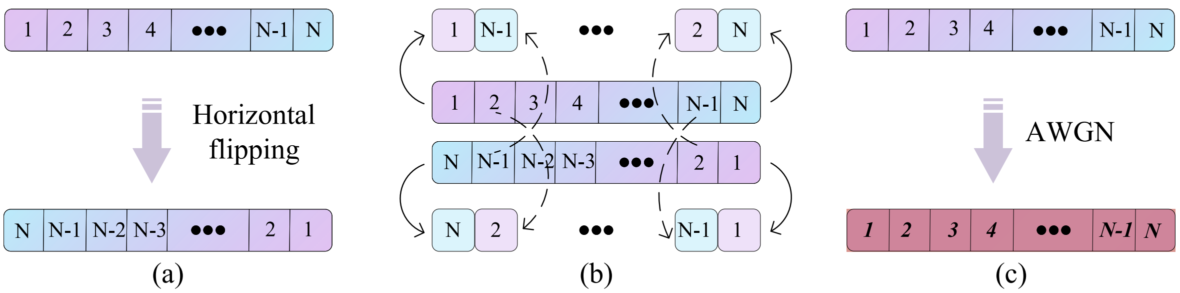


Fig. 5. Schematic diagrams of the proposed data augmentation method, where a square represents a symbol length. (a) Horizontal flipping; (b) cross-splicing; (c) addition of white Gaussian noise.

AMR methods use simulation signals in their research. The actual underwater acoustic channel environment is complex and dynamic, so it is difficult to simulate a real channel environment with high fidelity. Therefore, it is of great importance to design and test a deep learning-based AMR algorithm on real datasets. Since many of the existing datasets only have a limited amount of data, there is a compelling need for effective data augmentation schemes.

In actual communication systems, the transmitter usually uses the raised-cosine filter for pulse shaping of the baseband signal. Since the raised-cosine filter function is an even function, the waveform of a symbol has natural symmetry. Also, AMR does not involve the subsequent demodulation, so it is allowed to “destroy” the original signal to a certain extent. Because of this, this paper proposes a data augmentation method based on symmetry and microscale, and its implementation steps are shown in Fig. 5.

In Fig. 5, a square represents a symbol length, and the gradient of the square color indicate its relative position on the time axis. The proposed data augmentation method consists of three main steps: horizontal flipping, cross-splicing, and adding white Gaussian noise.

The specific steps are as follows:

Step 1: Horizontally flipping. The basic operation unit of this step is a sampling point. Assume a signal sequence with a length of n is expressed by $Y = [y_0, y_1, \ldots, y_{n-1}]$, and the signal sequence after horizontal inversion is given by $\widetilde{Y} = [y_{n-1}, y_{n-2}, \ldots, y_0]$. The amount of underwater acoustic signal data can be increased by one time through the horizontal flipping process;

Step 2: Cross-splicing. The basic operation unit of this step is a symbol. Assume a signal sequence is represented by $S = [s_0, s_1, \ldots, s_{m-1}]$, and the signal sequence after horizontal flipping be given by $\widetilde{S} = [\widetilde{s}_{m-1}, \widetilde{s}_{m-2}, \ldots, \widetilde{s}_0]$, where m is the number of symbols and is an even number, s_i and $\widetilde{s}_i$ represent a symbol before and after horizontal flipping, respectively. The elements with even subscripts in S and elements with odd subscripts in $\widetilde{S}$ are joined to obtain the reconstructed data $P = [s_0, \widetilde{s}_{m-1}, \ldots, s_{m-2}, \widetilde{s}_1]$. Similarly, elements with even subscripts in $\widetilde{S}$ and elements with odd subscripts in S are joined to obtain the reconstructed data $Q = [\widetilde{s}_{m-2}, s_1, \ldots, \widetilde{s}_0, s_{m-1}]$. The amount of underwater acoustic signal data can be increased by two times through the splicing and reconstruction processes;

Step 3: Adding white Gaussian noise. In this part, a signal with a certain SNR is obtained by introducing white Gaussian noise to the original signal and the signals obtained by the first two steps, which further increases the data volume by a factor of four.

By performing the above-mentioned three steps of the proposed data augmentation method, the dataset can be increased by seven times, which can be considered a significant improvement for data-driven deep learning-based algorithms. Using the enhanced dataset to train the model can reduce the overfitting of the network model to a certain extent and improve model robustness, as will be shown later in this paper.

V. Network Structure and Dataset

A. Network Structure and Complexity

To study the network structure’s influence on model performance, this paper combines the One2Three block with the dual- or single-stream SE block (the single-stream SE block is obtained by removing the SE block 2 from the dual-stream SE block) and designs four network models: One2ThreeNet-D22, One2ThreeNet-S22, One2ThreeNet-D14, and One2ThreeNet-S12. The naming rule for the models is as follows: One2ThreeNet - [dual- or single-stream] - [the number of SE blocks] - [the number of convolution layers in each SE block]. For instance, “D22” means that the model contains two dual-stream SE blocks, each of which contains two convolutional layers. The detailed structural parameters of the four network models are shown in Table I.

In this part, we use One2ThreeNet-D22 as an example to introduce the network structure of the One2ThreeNet series models. The structure of the One2ThreeNet-D22 model is shown in Fig. 6.

As shown in Fig. 6, the signal sequence with a length of 19,200 is preprocessed and then input into the One2Three block; a feature map with a dimension of 16 × 16 × 3 is obtained after performing the temporal feature extraction. Then, 16 convolutional filters with dimensions of 1 × 1 are used to expand the number of channels of the feature map to 16, and they are input into the first dual-stream SE block to obtain a 16 × 16 × 16 feature map. Similarly, 32 convolutional filters with dimensions of 1 × 1 expand the number of channels again and input them into the second dual-stream SE block to obtain a 16 × 16 × 32 feature map. To reduce the amount of network computation, the One2Threenet-D22

TABLE I

THE ONE2THREENET SERIES MODEL STRUCTURE

Layer name	Output size	One2ThreeNet-D22	One2ThreeNet-S22	One2ThreeNet-D12	One2ThreeNet-D14
One2Three	16 × 16 × 3	LSTM 1[unit of 16, input of 1200, output of 16] → feature map 1[16 × 16] LSTM 2[unit of 16, input of 600, output of 16] → feature map 2[16 × 16] LSTM 3[unit of 16, input of 300, output of 16] → feature map 3[16 × 16] Concatenate at the third dimension			
Expand_1	16×16×16	Conv [1 × 1 × 16], stride of 1, zero padding			
DualSE_1 or SingleSE_1	16 × 16 × 16	"Conv stream" Conv [3 × 3 × 16] Conv [3 × 3 × 16] Fc1 [16, 4] Fc2 [4, 16] r = 4 "Shortcut stream" Fc1 [16, 4] Fc2 [4, 16] r = 4	"Conv stream" Conv [3 × 3× 16] Conv [3 × 3 × 16] Fc1 [16, 4] Fc2 [4, 16] r = 4	"Conv stream" Conv [3 × 3 × 16] Conv [3 × 3 × 16] Fc1 [16, 4] Fc2 [4, 16] r = 4 "Shortcut stream" Fc1 [16, 4] Fc2 [4, 16] r = 4	"Conv stream" Conv [3 × 3 × 16] Conv [3 × 3 × 16] Conv [3 × 3 × 16] Conv [3 × 3 × 16] Fc1 [16, 4] Fc2 [4, 16] r = 4 "Shortcut stream" Fc1 [16, 4] Fc2 [4, 16] r = 4
Expand _2 or None	16 × 16 × 32 or 16 × 16 × 16	Conv [1 × 1 × 32], stride of 1, zero padding		—	
DualSE_2 or SingleSE_2	16 × 16 × 32 or 16 × 16 × 16	"Conv stream" Conv [3 × 3 × 32] Conv [3 × 3 × 32] Fc1[32, 8] Fc2 [8, 32] r = 4 "Shortcut stream" Fc1[32, 8] Fc2 [8, 32] r = 4	"Conv stream" Conv [3 × 3 × 32] Conv [3 × 3 × 32] Fc1[32, 8] Fc2 [8, 32] r = 4	—	—
Pool	8 × 8 × 32 or 8 × 8 × 16	Max Pool 2×2, stride 2			
Fc	1 × 128	Fc [2048, 128] Dropout of 0.5	Fc [2048, 128] Dropout of 0.5	Fc [1024, 128] Dropout of 0.5	Fc [1024, 128] Dropout of 0.5
Out	1 × 8	Fc [128, 8], Softmax function			

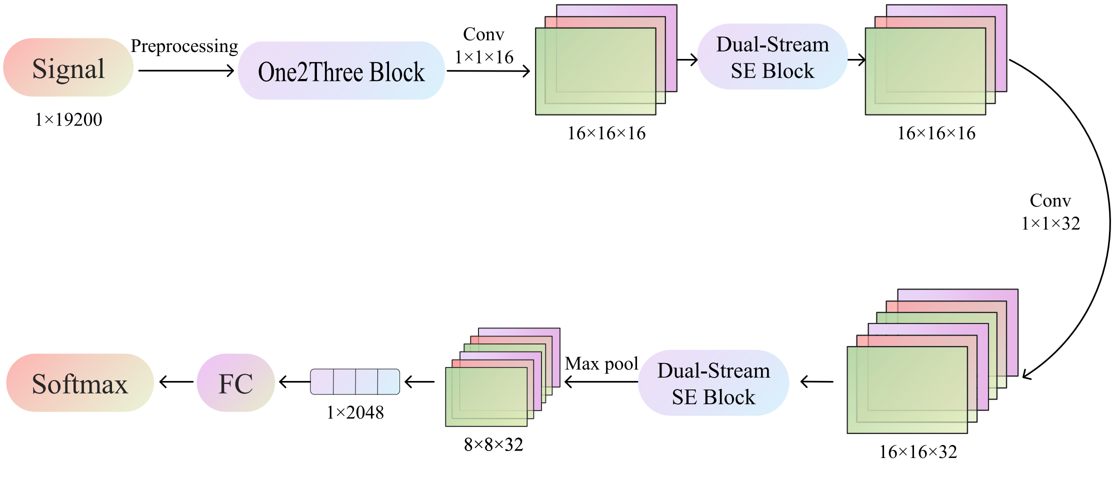


Fig. 6. The One2ThreeNet-D22 structure.

performs the 2 × 2 maximum pooling operation on the feature map to compress the feature map to the size of 8 × 8 × 32, and then the feature map is reconstructed to a 1 × 2,048 feature vector by the flattening layer. Finally, the prediction probabilities of different modulation modes are obtained by two fully-connected neural network layers and a layer with a SoftMax activation function.

To demonstrate the effectiveness of the One2ThreeNet, we selected some popular neural network models as baseline schemes, including MCLDNN [20], R&CNN [21], ResNet-12 [28], and SENet-12 [30]. Batch normalization (BN) was performed after all convolution operations to increase the stability of the model training process. In addition, the fully-connected layer in the SE block adopted the sigmoid activation function, while the other activation functions were the Relu function.

In this section, two indexes are used to evaluate the space and time complexity of each model, which are the number of parameters and FLOPs, respectively. The number of parameters of the convolution layer is as follows:

$$parameters_{conv} = (k_w \times k_h \times C_{in} + 1) \times C_{out}, \tag{9}$$

the FLOPs of the convolution layer is as follows:

$$FLOPs_{conv} = [(k_w \times k_h \times C_{in}) + (k_w \times k_h \times C_{in} - 1) \\ +1] \times C_{out} \times w \times h, \tag{10}$$

where $k_w \times k_h \times C_{in}$ represents the size of convolution kernel; C_{out} represents the number of convolution kernel, $w \times h$ represents the width and height of the feature map.

The number of parameters of the fully connected layer is as follows:

$$parameters_{fc} = (N_{in} + 1) \times N_{out}, \tag{11}$$

the FLOPs of the fully connected layer is as follows:

$$FLOPs_{fc} = [N_{in} + (N_{in} - 1) + 1] \times N_{out}, \tag{12}$$

where N_{in} represents the number of features in the input layer; N_{out} represents the number of features in the output layer.

The number of parameters of LSTM is as follows:

$$parameters_{LSTM} = 4 \times [(N_{embedding} + N_{hidden}) \\ \times N_{hidden} + N_{hidden}], \tag{13}$$

the FLOPs of LSTM can be approximated by the following equation:

$$FLOPs_{LSTM} = (N_{embedding} + N_{hidden}) \times N_{hidden} \times 4 \times 2, \tag{14}$$

where $N_{embedding}$ represents the embedding dimension; N_{hidden} represents the hidden state dimension.

The number of parameters and FLOPs of each model are shown in Table II. Due to the large size of the convolution kernel and the extremely large number of feature map channels, the number of parameters of R&CNN is much higher than those of other models, and the space complexity is relatively high. However, because of the very small number of convolution layers, the number of parameters of MCLDNN is only 0.08M. At the same time, One2ThreeNet only uses the convolution kernel with sizes of 3 × 3 and 1 × 1, and the number of channels in the feature map is small. Therefore, it also has extremely low spatial complexity. Since R&CNN uses a double-layer gated recurrent unit (GRU) structure with many hidden neurons, as well as a great number of large-sized convolution kernels, its FLOPs are significantly higher than other models. MCLDNN and One2ThreeNet-D12 benefit from fewer hidden layer neurons and a simple network structure, thus achieving extremely low time complexity.

TABLE II
The Number of Parameters and FLOPs for Each Model

Model	Parameters (M)	FLOPs (M)
One2ThreeNet-D22	0.44	16.99
One2ThreeNet-S22	0.43	16.99
One2ThreeNet-D12	0.28	7.02
One2ThreeNet-D14	0.29	9.38
R&CNN	9.79	504.87
MCLDNN	0.08	2.46
SENet-12	0.68	32.32
ResNet-12	0.67	32.30

B. Datasets

Due to the strong time-varying characteristics of underwater acoustic channels, accompanied by severe attenuation and strong noise interference, automatic modulation identification of a measured underwater acoustic signal is very challenging [31]. In this paper, two measured underwater acoustic signal datasets were used to analyze the performance of the proposed algorithm. The two datasets were collected at the Yellow Sea and the South China Sea. Among them, the Yellow Sea experimental dataset was collected in the offshore waters with bad sea conditions and serious Doppler effects, while the South China Sea experimental dataset was collected in the near shore waters with severe multipath effects and strong noise interference. These two datasets can well verify the robustness of the algorithm to various types of interferences. The parameter settings of the Yellow Sea dataset and the South China Sea dataset are provided in Tables III.

VI. Experimental Results and Analysis

In this paper, to ensure a fair comparison, the training hyperparameters of all models were set uniformly, the learning rate is 0.0005, the optimizer is Adam, and the batch size is 32. The experiments were performed on a personal computer equipped with an Intel(R) Core (TM) i5-9400 CPU @ 2.90GHz and, an NVIDIA GeForce GT 710 GPU, which operated on Windows 10 64bit. Python 3.7 and Pytorch 1.8.0 were used to implement the model. The data were divided into training, validation, and test sets according to the ratio of 6:2:2.

A. Microscale Rationality Verification

A single microscale control experiment was performed to verify the rationality of the three microscales. In this experiment, each network used an LSTM for time series feature extraction. The number of time steps and the number of hidden layer neurons were the same and equal to 16. After expanding the feature matrix into a one-dimensional vector, it was input to a fully-connected neural network with 64 hidden layer neurons to obtain the recognition result. The recognition accuracy results obtained at different microscales are presented in Table IV. This experiment used the South China Sea experimental dataset.

As shown in Table IV, the accuracy at the 0.25-symbol microscale was only 62.29%, while those at two-, single-, and half-symbol microscales were all higher than 90%, indicating that the proposed three types of microscales were reasonable.

TABLE III
THE EXPERIMENTAL PARAMETERS SETTING OF THE YELLOW SEA DATASET AND THE SOUTH CHINA SEA DATASET

Parameter	The Yellow Sea dataset	The South China Sea dataset
SNR	3.72 dB	-7.4 dB
Wind power	Grade 3	Grade 6
Wave height	0.5 m	3 m
Depth of the sea	3 m	80 m
DAC/ADC sampling rate	0.336 MHz	0.336 MHz
Carrier frequency	10 kHz	10 kHz
Symbol rate	560 Baud	560 Baud
Communication distance	7 m	1 km
Modulation type	2FSK, 4FSK, 2PSK, 4PSK, 16QAM, 64QAM, OFDM	DSSS, 2FSK, 4FSK, 2PSK, 4PSK, 16QAM, 64QAM, OFDM
Sample size	200 Samples for each type of modulation, and 1,400 Samples in total	600 Samples for each type of modulation, and 4,800 Samples in total

TABLE IV
COMPARISON OF THE RECOGNITION ACCURACY RESULTS OF A SINGLE-MICROSCALE EXPERIMENT FOR A SYMBOL LENGTH OF 600

	4 Symbols	2 Symbols	1 Symbol	0.5 Symbols	0.25 Symbols
LSTM	[16, 2,400, 16]	[16, 1200, 16]	[16, 600, 16]	[16, 300, 16]	[16, 150, 16]
Fc	Fc [256, 64], dropout of 0.5				
Out	Fc [64, 8], softmax function				
Accuracy (%)	84.48	90.21	90.83	90.83	62.29

In addition, an accuracy of 84.48% has been achieved at the four-symbol microscale, which showed that the microscale selection was upwardly compatible. Even when there was a certain difference in the code rate between the signals to be identified, it was feasible to extract signal features at a fixed microscale.

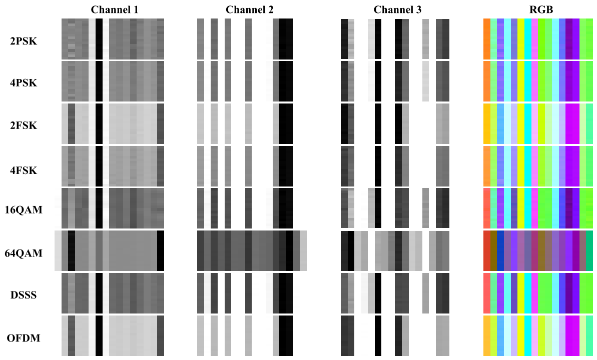


Fig. 7. The comparison of feature maps extracted by the One2Three block from the South China Sea experiment dataset.

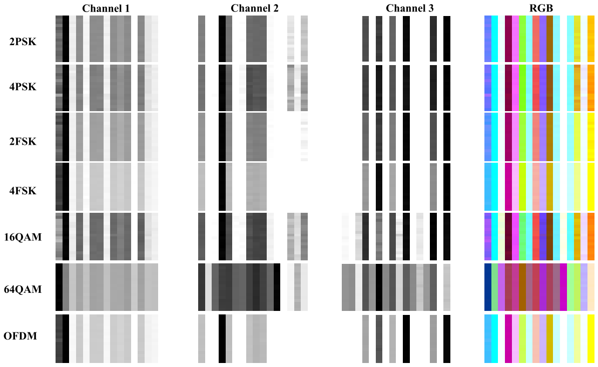


Fig. 8. The comparison of feature maps extracted by the One2Three block from the Yellow Sea experiment dataset.

B. One2Three Block Performance Analysis

Excellent features should have sufficient discrimination and robustness. To show the discrimination of the feature maps extracted by the proposed One2Three block, this paper uses the One2ThreeNet-D22 model to perform several visual analyses on feature maps obtained from different types of signals. The feature maps of the underwater acoustic signals with eight different modulation modes from the two datasets are presented in Figs. 7 and 8. We map the values of the feature maps of the three channels obtained by the One2Three block to the corresponding grayscale images. At the same time, we regard these three groups of grayscale images as three channels of RGB images, and give the color map after the superposition of the three channels, To see the difference between the feature maps more intuitively. It can be seen that these feature maps exhibit a regular and interesting barcode-like trend. There were significant differences between the feature maps of the underwater acoustic signals with different types of modulation modes, which were visually reflected in the color shading in different areas of the feature map. This means that the feature maps extracted by the One2Three block were very distinguishable, and they could characterize the intrinsic characteristics of different modulation modes. In addition, for modulation modes of different orders from the same category, differences in some channel signatures were significantly reduced, but some channels still had enough differentiated regions, which indicates that features extracted at three microscales were complementary. Moreover, differences between signals of different orders in the same category were small, which was consistent with the existing research results.

In addition, to illustrate the robustness of the feature map obtained by the One2Three block, eight samples of the 2PSK and OFDM signals were randomly selected from the South China Sea experiment dataset to visualize the feature

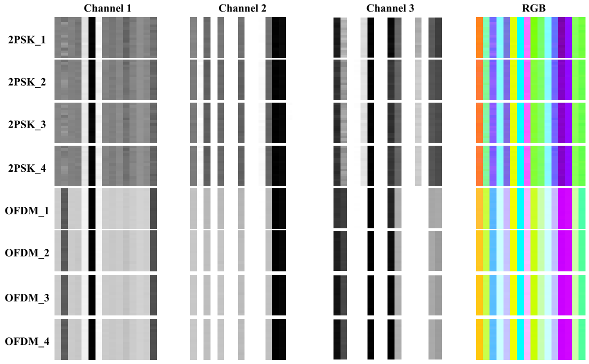


Fig. 9. The comparison of feature maps extracted by the One2Three block from the South China Sea experiment dataset.

map. As shown in Fig. 9, the feature maps of different samples of the same modulation showed a high degree of consistency, which indicated that the feature map obtained by the One2Three block automatically filtered the random noise among different samples and successfully captured the commonality between different samples of the same modulation.

C. One2ThreeNet Performance Analysis

The One2ThreeNet-D22 model was selected to compare the recognition performance of the One2ThreeNet series model and the baseline model on different datasets to illustrate the superiority of the One2ThreeNet series model. The channel weights of feature maps were calculated to illustrate the effectiveness of the dual-stream SE block module. Finally, the performance evaluation indexes of the models were compared on different datasets to analyze the models quantitatively. Based on the Yellow Sea experimental dataset and the South China Sea experimental dataset, One2ThreeNet-D22 is compared with the training loss curves of other baseline models as shown in Fig. 10 and Fig. 11.

As shown in Fig. 10 and Fig. 11, among all models, the One2ThreeNet-D22 model showed the fastest decline in the training loss curve, followed by the R&CNN, MCLDNN, SENet-12, and ResNet-12 models. Neither of the latter four models could reduce the loss value to a lower level than One2ThreeNet-D22. The results indicated that the One2Three block module could effectively extract the temporal features of the underwater acoustic signal, while the subsequent CNN module, connected in series, could effectively extract the spatial features of the signal. R&CNN can also obtain a lower loss value due to the first temporal features extraction and then spatial features extraction, but the curve fluctuates greatly, which may be due to the complex network structure. Although the MCLDNN model could also recognize modulations by combining spatial and temporal signal features, it is not recommended to extract spatial features first by a CNN and then extract temporal features by an LSTM because the temporal features of one-dimensional signals are more obvious. However, the SENet-12 and ResNet-12 models, which are too monotonous for modulation recognition, could extract only spatial features.

To illustrate the excitation and suppression performances of the dual-stream SE block on different channel profiles, the average activation weight of the second dual-stream SE block module obtained by the One2ThreeNet-D22 model is shown in Fig. 12. In the Conv-Stream block, the weights of different channels varied significantly, indicating that the network could learn the differences between different channels of the feature map and select them. In the Shortcut-Stream block, there were certain differences between different channels, but they were relatively small. This shows that the two convolution layers in the Conv-Stream block could extract the feature map at a high level, enhancing differences between different channels. So, based on the results, it was necessary to perform SE operations on residual channels.

To analyze the performance of each model quantitatively, the South China Sea dataset and the Yellow Sea dataset were used to conduct 20 repetitive experiments using the One2ThreeNet series models and baseline models. The results are shown in Tables V and VI.

As shown in Tables V and VI, the One2ThreeNet-D22 model achieved the best results in terms of accuracy, loss value, and recognition time among all models and had a very high recognition performance and efficiency. This indicated that the One2ThreeNet-D22 model could learn well the relationship between the signal characteristics and the corresponding modulation modes without consuming much computing resources. At the same time, the MCLDNN model achieved recognition accuracy of more than 93%, which indicated that the combination of spatial and temporal features was feasible for modulation recognition. However, because the temporal features of one-dimensional signals were more obvious than spatial features, it was not effective to extract spatial features first and then temporal features. The SENet-12 and ResNet-12 models had the worst recognition performance among all models because they extracted only spatial features and their feature sets were too monotonic.

The recognition performance comparison results of the One2ThreeNet series models on the Yellow Sea and South China Sea experimental datasets are presented in Tables VII and VIII, respectively.

The results show that the One2ThreeNet series models achieve a recognition accuracy of more than 99% and each index has a small gap, showing very high recognition accuracy and high stability. This indicates that the One2Three block could successfully convert the temporal features of the signal into a feature map so that subsequent CNN modules could accurately extract the spatial features of the feature map and realize AMR. In addition, among all models, the One2ThreeNet-D22 model achieves the best performance, which demonstrates that the dual-stream SE block's double SE design could optimize the channel expression of the signature map, thus effectively extracting spatial features.

D. Performance Analysis With Noise Data

In this section, we analyzed the robustness of each model to noise by comparing the recognition accuracy of the trained model on the test set with Gaussian noise. The test sets part of the South China Sea experimental dataset was selected for the signal to be identified, and the SNR was between 10-30dB. We calculated the average recognition accuracy (unit: %) of

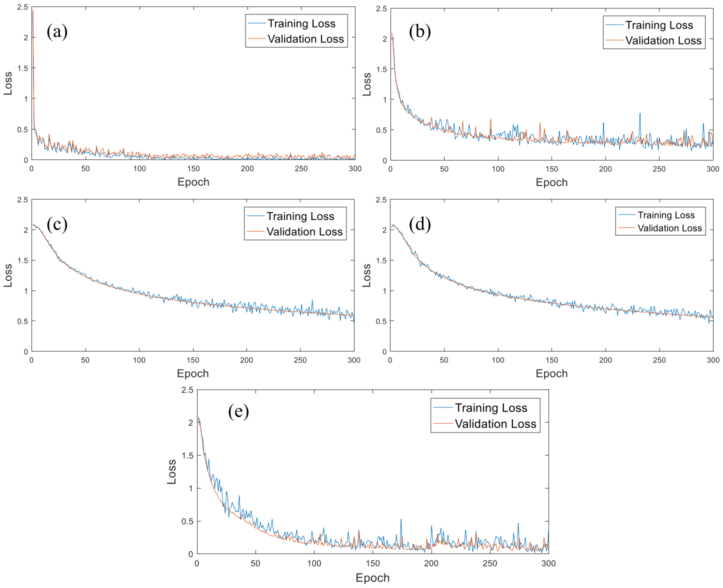


Fig. 10. The training loss curves of the One2ThreeNet-D22 model and baseline models on the South China Sea experimental dataset. (a): One2ThreeNet-D22, (b): MCLDNN, (c): SENet-12, (d): ResNet-12, (e): R&CNN.

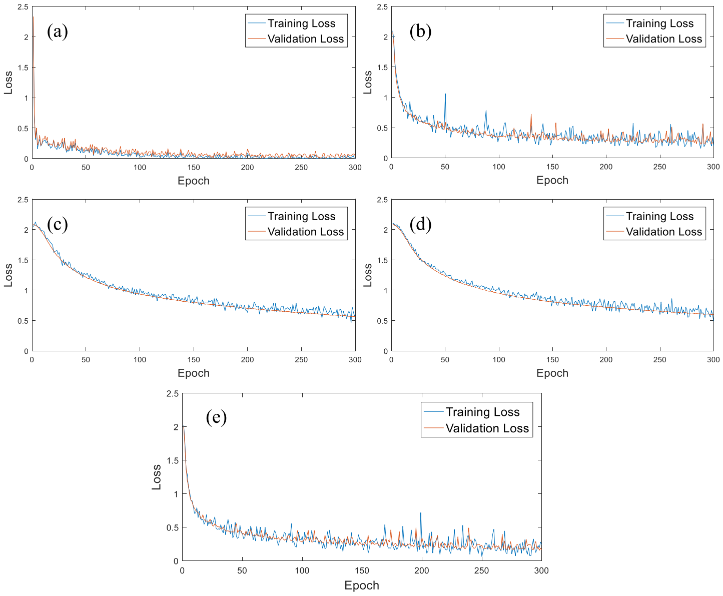


Fig. 11. The training loss curves of the One2ThreeNet-D22 model and baseline models on the Yellow Sea experimental dataset. (a): One2ThreeNet-D22, (b): MCLDNN, (c): SENet-12, (d): ResNet-12, (e): R&CNN.

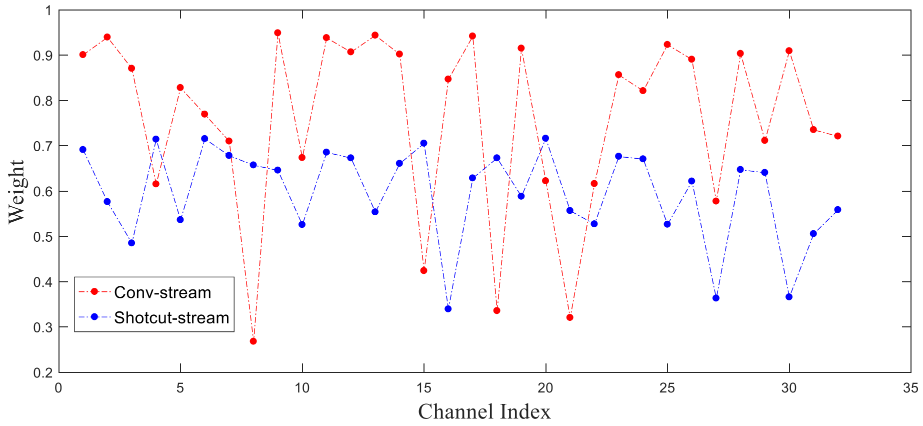


Fig. 12. The average activation weights of the dual-stream SE block and the One2ThreeNet-D22 based on the South China Sea experimental dataset.

20 independent repeated experiments, the results are shown in Table IX.

As shown in Table IX, the performance of One2ThreeNet-D22 and R&CNN is almost the same under high SNR, but at

TABLE V

PERFORMANCE COMPARISON OF THE ONE2THREENET-D22 MODEL AND BASELINE MODELS ON THE SOUTH CHINA SEA EXPERIMENTAL DATASET

	One2ThreeNet-D22	R&CNN	MCLDNN	SENet-12	ResNet-12
Optimal accuracy on the test set (%)	99.688	99.688	93.646	88.854	88.125
Average accuracy on the test set (%)	99.427	99.214	91.074	88.169	88.003
Optimal loss on the validation set	0.0036	0.0363	0.2229	0.5625	0.5927
Optimal loss on the training set	0.0004	0.0163	0.1532	0.4484	0.4599

TABLE VI

PERFORMANCE COMPARISON OF THE ONE2THREENET-D22 MODEL AND BASELINE MODELS ON THE YELLOW SEA DATASET

	One2ThreeNet-D22	R&CNN	MCLDNN	SENet-12	ResNet-12
Optimal accuracy on the test set (%)	99.643	99.286	92.857	87.857	88.571
Average accuracy on the test set (%)	99.335	98.864	91.249	87.524	88.404
Optimal loss on the validation set	0.0034	0.1582	0.2283	0.5641	0.5944
Optimal loss on the training set	0.0003	0.0658	0.1592	0.4662	0.4880

TABLE VII

PERFORMANCE COMPARISON OF THE ONE2THREENET SERIES MODELS ON THE SOUTH CHINA SEA DATASET

	One2ThreeNet-D22	One2ThreeNet-D14	One2ThreeNet-D12	One2ThreeNet-S22
Optimal accuracy on the test set (%)	99.688	99.583	99.583	99.061
Average accuracy on the test set (%)	99.427	99.476	99.369	98.707
Optimal loss on the validation set	0.0036	0.0066	0.0094	0.0302
Optimal loss on the training set	0.0004	0.0034	0.0077	0.0126

TABLE VIII

PERFORMANCE COMPARISON OF THE ONE2THREENET SERIES MODELS ON THE YELLOW SEA EXPERIMENTAL DATASET

	One2ThreeNet-D22	One2ThreeNet-D14	One2ThreeNet-D12	One2ThreeNet-S22
Optimal accuracy on the test set (%)	99.643	99.286	99.286	99.286
Average accuracy on the test set (%)	99.335	99.146	99.201	98.811
Optimal loss on the validation set	0.0034	0.0087	0.0144	0.0109
Optimal loss on the training set	0.0003	0.0061	0.0051	0.0094

TABLE IX

PERFORMANCE COMPARISON OF THE ONE2THREENET-D22 MODEL AND BASELINE MODELS ON THE SOUTH CHINA SEA EXPERIMENTAL DATASET

	One2ThreeNet-D22	R&CNN	MCLDNN	SENet-12	ResNet-12
30 dB	99.624	99.382	92.461	88.154	88.162
25 dB	94.152	93.127	86.462	86.675	84.134
20 dB	78.753	72.664	62.766	53.412	53.522
15 dB	68.422	50.152	24.331	12.534	12.566
10 dB	36.517	25.688	12.513	12.586	12.524

TABLE X

PERFORMANCE COMPARISON OF THE ONE2THREENET-D22 MODEL FOR DIFFERENT DATA AUGMENTATION ALGORITHMS

	Original	Data augmentation
Optimal accuracy on the test set (%)	74.286	86.071
Average accuracy on the test set (%)	72.694	85.104

low SNR, One2ThreeNet-D22 is slightly better than R&CNN, which may be due to the over fitting caused by the complex network structure of R&CNN. However, when the SNR is reduced to less than 20dB, the accuracies of MCLDNN, SENet-12, and ResNet-12 have all dropped significantly, which show that these models are difficult to effectively extract the features of low SNR signals. Therefore, the structure of extracting temporal features first and then spatial features makes One2ThreeNet-D22 and R&CNN more robust to noise than other models.

E. Performance Analysis of Data Augmentation Methods

To verify the performance of data augmentation methods, the One2ThreeNet-D22 model was tested on the Yellow Sea experimental dataset, one-seventh of the training sets and validation sets were uniformly selected to study the robustness gains brought by the data augmentation methods. In this part, we use the proposed augmentation algorithm to expand the training sets by seven times, before data augmentation, the size of the training sets, validation sets and test sets were 120, 40, and 280, respectively. After data augmentation, the size of the training sets, validation sets, and test sets are 960, 320, and 280, respectively. The results are shown in Table X.

As shown in Table X, the recognition performance indicators' values before and after data augmentation differed significantly. After data augmentation, the sample size was expanded by seven times, the optimal recognition accuracy was increased from 74.286% to 86.071%, and the average accuracy was also increased from 72.694% to 85.104%, indicating that the data augmentation algorithm proposed in this paper can significantly improve the richness of the dataset, and thus improve the accuracy and robustness of the model.

VII. Conclusion

To address the problem of the high cost of underwater acoustic signal acquisition, this paper proposed a data augmentation method for AMR, which can increase the sample size by seven times and solve the problem of lacking sufficient data for deep learning-based AMR methods. In addition, aiming at the problem that the existing methods cannot extract the temporal features of underwater acoustic signals effectively, the concept of microscale was proposed, which converts underwater acoustic signals into time series and provides stronger interpretability. On this basis, a temporal feature extractor named the One2Three block, which can automatically extract the temporal features of signals and convert them into three-channel feature maps, was proposed. Finally, this paper proposed a spatial feature extractor named the dual-stream SE block, which can abstract and synthesize the feature map transformed by the One2Three block from the perspective of spatial features to realize AMR. The proposed method has very low spatial complexity and time complexity and is highly robust to noisy data. The experimental results on the underwater acoustic signal datasets of the Yellow Sea experiment and the South China Sea experiment showed that the proposed data augmentation algorithm can effectively improve model stability. The feature map obtained by the One2Three block has high discrimination and robustness. In addition, by using the proposed method, the DSSS, 2FSK, 4FSK, 2PSK, 4PSK, 16QAM, 64QAM, and OFDM signals can be accurately identified in the offshore waters where the sea conditions are bad, and the Doppler effect is severe, as well as in the near shore waters where the multipath effect and noise interference are severe. However, it has not been analyzed whether the proposed method can handle extremely severe deep seas. In future work, we plan to conduct additional underwater acoustic communication experiments in different sea areas to explore and analyze the shortcomings of the proposed algorithms further.

References

[1] L. Wan, H. Zhou, and X. Xu, “Adaptive modulation and coding for underwater acoustic OFDM,” *IEEE J. Ocean. Eng.*, vol. 40, no. 2, pp. 327–336, Apr. 2015.

[2] H. C. Wu, M. Saquib, and Z. Yun, “Novel automatic modulation classification using cumulant features for communications via multipath channels,” *IEEE Trans. Wireless Commun.*, vol. 7, no. 8, pp. 3098–3105, Aug. 2008.

[3] Y. Wang, G. Gui, T. Ohtsuki, and F. Adachi, “Multi-task learning for generalized automatic modulation classification under non-Gaussian noise with varying SNR conditions,” *IEEE Trans. Wireless Commun.*, vol. 20, no. 6, pp. 3587–3596, Jun. 2021.

[4] Z. Huang, S. Li, X. Yang, and J. Wang, “OAE-EEKNN: An accurate and efficient automatic modulation recognition method for underwater acoustic signals,” *IEEE Signal Process. Lett.*, vol. 29, pp. 518–522, 2022.

[5] Z. Lei et al., “Towards recurrent neural network with multi-path feature fusion for signal modulation recognition,” *Wireless Netw.*, vol. 28, no. 2, pp. 551–565, Feb. 2022.

[6] A. K. Ali and E. Erçelebi, “Automatic modulation recognition of DVB-S2X standard-specific with an APSK-based neural network classifier,” *Measurement*, vol. 151, Feb. 2020, Art. no. 107257.

[7] O. Russakovsky et al., “ImageNet large scale visual recognition challenge,” *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[8] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, “A survey of deep neural network architectures and their applications,” *Neurocomputing*, vol. 234, pp. 11–26, Apr. 2017.

[9] Z. Zhang, Y. Zou, and C. Gan, “Textual sentiment analysis via three different attention convolutional neural networks and cross-modality consistent regression,” *Neurocomputing*, vol. 275, pp. 1407–1415, Jan. 2018.

[10] C. Luo, J. Ji, Q. Wang, X. Chen, and P. Li, “Channel state information prediction for 5G wireless communications: A deep learning approach,” *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 1, pp. 227–236, Jan./Mar. 2020.

[11] K. Bu, Y. He, X. Jing, and J. Han, “Adversarial transfer learning for deep learning based automatic modulation classification,” *IEEE Signal Process. Lett.*, vol. 27, pp. 880–884, 2020.

[12] Z. Pan, S. Wang, M. Zhu, and Y. Li, “Automatic waveform recognition of overlapping LPI radar signals based on multi-instance multi-label learning,” *IEEE Signal Process. Lett.*, vol. 27, pp. 1275–1279, 2020.

[13] Y. Tu, Y. Lin, C. Hou, and S. Mao, “Complex-valued networks for automatic modulation classification,” *IEEE Trans. Veh. Technol.*, vol. 69, no. 9, pp. 10085–10089, Sep. 2020.

[14] M. Marey and H. Mostafa, “Soft-information assisted modulation recognition for reconfigurable radios,” *IEEE Wireless Commun. Lett.*, vol. 10, no. 4, pp. 745–749, Apr. 2021.

[15] T. J. O’Shea, J. Corgan, and T. C. Clancy, “Convolutional radio modulation recognition networks,” in *Proc. 17th Int. Conf. Eng. Appl. Neural Netw.*, San Juan, Puerto Rico, May 2016, pp. 213–226.

[16] T. J. O’Shea, T. Roy, and T. C. Clancy, “Over-the-air deep learning based radio signal classification,” *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 1, pp. 168–179, Feb. 2018.

[17] D. Wang et al., “Modulation format recognition and OSNR estimation using CNN-based deep learning,” *IEEE Photon. Technol. Lett.*, vol. 29, no. 19, pp. 1667–1670, Oct. 1, 2017.

[18] R. Li, L. Li, S. Yang, and S. Li, “Robust automated VHF modulation recognition based on deep convolutional neural networks,” *IEEE Commun. Lett.*, vol. 22, no. 5, pp. 946–949, May 2018.

[19] Z. Zhang, H. Luo, C. Wang, C. Gan, and Y. Xiang, “Automatic modulation classification using CNN-LSTM based dual-stream structure,” *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 13521–13531, Nov. 2020.

[20] J. Xu, C. Luo, G. Parr, and Y. Luo, “A spatiotemporal multi-channel learning framework for automatic modulation recognition,” *IEEE Wireless Commun. Lett.*, vol. 9, no. 10, pp. 1629–1632, Oct. 2020.

[21] W. Zhang, X. Yang, C. Leng, J. Wang, and S. Mao, “Modulation recognition of underwater acoustic signals using deep hybrid neural networks,” *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 5977–5988, Aug. 2022.

[22] X. Ji, J. Wang, Y. Li, Q. Sun, S. Jin, and T. Q. S. Quek, “Data-limited modulation classification with a CVAE-enhanced learning model,” *IEEE Commun. Lett.*, vol. 24, no. 10, pp. 2191–2195, Oct. 2020.

[23] M. Patel, X. Wang, and S. Mao, “Data augmentation with conditional GAN for automatic modulation classification,” in *Proc. 2nd ACM Workshop Wireless Secur. Mach. Learn.*, Linz, Austria, Jul. 2020, pp. 31–36.

[24] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[25] G. Van Houdt, C. Mosquera, and G. Nápoles, “A review on the long short-term memory model,” *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5929–5955, May 2020.

[26] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, 2015, pp. 1–14.

[27] C. Szegedy et al., “Going deeper with convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 1–9.

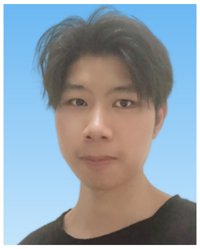
[28] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2016, pp. 770–778.

[29] F. Wang et al., “Residual attention network for image classification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6450–6458.

[30] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, “Squeeze-and-excitation networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[31] H. Khan, S. A. Hassan, and H. Jung, “On underwater wireless sensor networks routing protocols: A review,” *IEEE Sensors J.*, vol. 20, no. 18, pp. 10371–10386, Sep. 2020.

Jingjing Wang (Member, IEEE) received the B.S. degree in industrial automation from Shandong University, Jinan, China, in 1997, the M.Sc. degree in control theory and control engineering from Qingdao University of Science and Technology, Qingdao, China, in 2002, and the Ph.D. degree in computer application technology from the Ocean University of China, Qingdao, in 2012. From 2014 to 2015, she was a Visiting Professor with The University of British Columbia. She is currently a Professor with the School of Information Science and Technology, Qingdao University of Science and Technology. Her research interests include underwater wireless sensor networks, acoustic communications, ultrawideband radio systems, and MIMO wireless communications.



Zihao Huang received the B.S. degree in electronic information engineering from Nanchang Hangkong University, Nanchang, China, in 2018, and the M.Sc. degree in computer technology from Qingdao University of Science and Technology, Qingdao, China, in 2022. He is currently pursuing the Ph.D. degree with Xiamen University, Xiamen, China. His research interests include underwater acoustic communications, deep learning, and signal processing.

Wei Shi (Member, IEEE) received the B.S. degree from Ludong University, Yantai, China, in 2009, and the master's degree in signal and information processing and the Ph.D. degree in computer application technology from the Ocean University of China, Qingdao, China, in 2011 and 2014, respectively. He is currently an Associate Professor with the School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao. His research interests include 5G, 60GHz wireless communication, and underwater wireless sensor networks.

Shiwen Mao (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Polytechnic University, Brooklyn, NY, USA, in 2004. Currently, he is a Professor and the Earle C. Williams Eminent Scholar Chair of Electrical and Computer Engineering with Auburn University, Auburn, AL, USA. His research interest include wireless networks, multimedia communications, and smart grid. He is on the Editorial Board of IEEE/CIC China Communications, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE INTERNET OF THINGS JOURNAL, IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY, IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING, IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING, IEEE TRANSACTIONS ON MOBILE COMPUTING, IEEE MULTIMEDIA, IEEE NETWORKING LETTERS, *IEEE Network* magazine, and *ACM GetMobile*. He is a co-recipient of the 2021 IEEE Communications Society Outstanding Paper Award, the IEEE Vehicular Technology Society 2020 Jack Neubauer Memorial Award, the IEEE ComSoc MMTC 2018 Best Journal Paper Award and the 2017 Best Conference Paper Award, the Best Demo Award from IEEE SECON 2017, the Best Paper Awards from IEEE GLOBECOM 2019, 2016, and 2015, IEEE WCNC 2015, and IEEE ICC 2013, and the 2004 IEEE Communications Society Leonard G. Abraham Prize in the Field of Communications Systems. He is a member of the ACM.