Received December 9, 2020, accepted December 16, 2020, date of publication December 21, 2020,
date of current version December 31, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3046124

Automatic Modulation Classification for Short Burst Underwater Acoustic Communication Signals Based on Hybrid Neural Networks

YONGBIN LI, BIN WANG, GAOPING SHAO, AND SHUAI SHAO
PLA Strategic Support Force Information Engineering University, Zhengzhou 450001, China
Corresponding author: Bin Wang (oceansgroup2020@163.com)

ABSTRACT Automatic modulation classification (AMC) is challenging for short burst underwater acoustic (UWA) communication signals. Difficulties include but are not limited to the poor UWA channels, impulsive noise, and data scarcity. To address these problems, a method based on hybrid neural networks (HNNs) is proposed in this paper. First, an impulsive noise preprocessor is adopted to mitigate the impulse in the received signals. Subsequently, an HNN consisting of an attention aided convolutional neural network (Att-CNN) and a sparse auto-encoder is built to extract features from the temporal waveforms and square spectra of the preprocessed signals after burst detection. Finally, a late fusion is made to combine the prediction results of the two sub-networks. To overcome the variable signal duration relative to the fixed input size of the Att-CNN, a data-reusing approach is proposed to perform dimension preprocessing on the waveforms. Moreover, a transfer learning strategy is introduced to resolve the issue of insufficient training data from the testing channel. The results of simulation experiments and practical signal tests both demonstrate that the proposed method is robust against UWA channels and ambient noise. Our approach significantly outperforms existing deep learning-based methods in dealing with short and weak signal bursts, while requiring less training data from the testing channel.

INDEX TERMS Automatic modulation classification, hybrid neural networks, attention aided convolutional neural networks, sparse auto-encoder, data-reusing, transfer learning.

I. INTRODUCTION

Automatic modulation classification (AMC) plays an important role in the attribute identification and information recovery of received communication signals. In recent years, with the continuous development in ocean-related technologies and the growing demand for marine information acquisition, AMC for underwater acoustic (UWA) communication signals has emerged as an increasingly important research topic. However, because of the complexity of the marine environment, progress in this field has been slow. Especially in military applications, the transmitted signals are always short and burst, which has increased the difficulty of AMC.

Conventional AMC for UWA communication signals is mostly based on pattern recognition approaches. This is done in two steps: feature extraction and classification. Different features are first built on the basis of domain knowledge and are then fed to different classifiers for classification. Common features include instantaneous features [1], [2], cyclostationary features [3]–[7], high-order cumulants-based features [8]–[10], spectral features [11], [12], wavelet transform-based features [2], [9], [13], and time–frequency transform-based features [14], [15]. Generally, these hand-crafted features are not robust against the complexity of the marine environment, such as the poor UWA channels and impulsive noise. The performance of such methods relies on manual experience, and can only be optimum under certain conditions.

To reduce the dependency on domain knowledge and to extract more effective and stable features, many deep learning (DL)-based methods have been developed recently. Herein, we review the recent success achieved by DL in AMC for

The associate editor coordinating the review of this manuscript and approving it for publication was Jiajia Jiang.

TABLE 1. List of the recent success achieved by DL in AMC for UWA communication signals.

Network input	Network	Paper	Evaluation
IQ raw waveforms	CNN CNN + LSTM	[16] [17]	Requires an accurate estimation about signal carrier frequency
Signal waveforms on carrier	CNN DAE + SAE	[18] [19]	The classification categories are limited; Not robust against the impulsive noise environments
Instantaneous features of signal waveforms	LSTM	[20]	Requires a large amount of training data from the testing channel
Spectral sequence or diagrams of signals	SAE CNN	[21] [22]	Requires a sufficient number of transmitted symbols; performance deteriorates sharply when the channel fading is poor
Signal spectrograms	CNN	[23]	Requires a trade-off between the time and frequency resolutions; Unable to perform inter-class recognition for PSK signals

UWA communication signals, as summarized in Table 1. For each approach, different modality information of signals is taken as the input of different network. Moreover, an evaluation is presented in terms of classification categories, robustness against environments, and application limitations.

As shown in Table 1, Marcoux *et al.* [16] and Li-Da *et al.* [17] utilized the baseband IQ raw waveforms to train a convolutional neural network (CNN) and a hybrid network comprising a CNN sub-network and a long short-term memory (LSTM) sub-network, respectively. These methods are based on the assumption that the signal carrier frequencies are accurately estimated. However, this remains challenging under the effect of time-varying and multi-path fading UWA channels. Zhou *et al.* [18] trained a similar CNN with the real and imaginary parts of received signals, achieving a high classification accuracy for three types of practical UWA communication signals. Yang *et al.* [19] continued to improve the classification performance at low signal-to-noise ratios (SNRs). They used signal waveforms enhanced by a denoising auto-encoder (DAE) to train another sparse auto-encoder (SAE), and the performance under Gaussian noise was improved. However, the noise in marine environments is much more complex, thus requiring further feasibility testing. Yu *et al.* [20] proposed an LSTM network and fed it with instantaneous features of signal waveforms. The effectiveness of the algorithm was proven on practical UWA communication signals. However, a large amount of training data from the testing channel is required, since an LSTM has significantly more parameters than common networks. Moreover, for the aforementioned methods based on temporal waveforms and CNN, the variable signal duration is inconsistent with the fixed input size of the network.

Jiang *et al.* [21] and Li *et al.* [22] respectively trained an SAE and a CNN with the power spectra of received signals and the spectra obtained after square or quartic transformation. Their approaches could effectively recognize most common UWA communication signals. However, an accurate estimation of the power spectra requires a sufficient number of transmitted symbols. Thus, these methods may not apply to short burst UWA communication signals. Moreover, their performance deteriorates sharply when the channel fading is poor. Yao *et al.* [23] designed a deeper and wider CNN, which takes the spectrograms of the signals as input. This method is more robust against multi-path channels but leads to a trade-off between the time and frequency resolutions. As signal parameters vary, it is difficult to select an appropriate fast Fourier transform window length to ensure the quality of the spectrograms. Moreover, because the spectrograms only contain the amplitude information of signals, this method is unable to perform inter-class recognition for phase shift keying (PSK) signals.

To improve the recognition performance for short burst UWA communication signals in complex marine environments, the idea of multimodal DL [24]–[28] is introduced and a novel hybrid neural networks (HNNs)-based AMC method is proposed in this paper. First, impulsive noise preprocessing, burst detection, and dimension preprocessing are performed on the received noisy signals. Subsequently, the temporal waveforms of the dimension-preprocessed signals are fed to an attention-aided convolutional neural network (Att-CNN) for preliminary classification. If a signal is recognized as PSK-modulated, another SAE will be further adopted to extract features from the signal square spectra. Finally, a late fusion is implemented to combine the results of the two modalities (i.e., temporal waveforms and square spectra).

Moreover, we adopt the idea of transfer learning and build a transfer data model to overcome the issue of insufficient training data from the testing channel. The results of simulation experiments and practical signal tests both demonstrate that the proposed method is robust against UWA channels and ambient noise. It can effectively recognize common UWA communication signals including PSK, frequency shift keying (FSK), orthogonal frequency division multiplexing (OFDM), and sweep spread carrier (S2C) [29] signals. Our method is also proven to have better performance in dealing with short and weak signal bursts, while requiring less training data from the testing channel.

This paper introduces an innovative approach for the AMC of short burst UWA communication signals. The advantages and contributions of our work are summarized as follows:

- Most conventional algorithms rely on domain knowledge, and hand-crafted features cannot be effectively generalized to different underwater environments. In comparison, the proposed method adopts deep neural networks for automatic feature extraction and classification. Thus, the performance is more stable.
- A self-attention mechanism is introduced into the proposed HNN to help extract more effective signal

features. Thus, weak signals can be better recognized. By contrast, the performance of existing methods declines at low SNRs.

- Most existing DL-based methods require the dimension of testing signals to be the same as that of the network input. However, our approach can handle signals with variable duration via the proposed dimension preprocessing technique of data-reusing.
- Existing DL-based methods are trained on large amounts of data obtained from a testing channel. In comparison, the proposed method significantly reduces this requirement via introducing the transfer learning strategy and the presented transfer data model.

The remainder of this paper is organized as follows: Sec. II introduces the system model and proposed method. Sec. III presents the experimental results and discussion. Finally, Sec. IV concludes this paper.

To foster manuscript readability, Table 2 and Table 3 summarizes the acronyms and mathematical notions used in this paper, respectively.

TABLE 2. List of the acronyms used in this paper.

Acronym	Definition
AMC	Automatic modulation classification
UWA	Underwater acoustic
DL	Deep learning
CNN	Convolutional neural network
LSTM	Long short-term memory
SNR	Signal-to-noise ratio
SAE	Sparse auto-encoder
HNN	Hybrid neural network
Att-CNN	Attention-aided convolutional neural network
PSK	Phase shift keying
FSK	Frequency shift keying
OFDM	Orthogonal frequency division multiplexing
S2C	Sweep spread carrier
MSNR	Mixed signal-to-noise ratio
INP	Impulsive noise preprocessor
SDGAN	Signal denoising generative adversarial network
DR	Data-reusing
FC	Fully connected
1D Conv	One-dimensional convolutional
SE	Squeeze and excitation
TTC	Target testing channel
ZP	Zero-padding
STFT	Short-time Fourier transformation

II. SYSTEM MODEL AND PROPOSED METHOD

In this section, we describe the proposed method, starting from the signal model construction in Sec. II.A. We then focus, in Sec. II.B, on the proposed AMC model shown in Fig. 1, which comprises a preprocessing part and a classification part. The preprocessing part includes impulsive noise preprocessing, burst detection, and dimension preprocessing. The classification algorithm is performed with an Att-CNN and a SAE module, which are further fused with a late fusion

TABLE 3. List of the mathematical notions used in this paper.

Symbol	Definition
$y(n)$	Received signal
$s(n)$	Transmitted signal
$h(n)$	Impulse response of the UWA channel
$w(n)$	Ambient noise
L	Sample number of $y(n)$
$\phi(u)$	Characteristic function of alpha-stable distributed noise
α	Characteristic exponent of alpha-stable distributed noise
a	Location parameter of alpha-stable distributed noise
β	Symmetry parameter of alpha-stable distributed noise
γ	Dispersion of alpha-stable distributed noise
σ_s^2	Variance of the signal
$I(n)$	Impulsive noise preprocessed signal
$b(n)$	Output after burst detection
L_b	Sample number of $b(n)$
$d_i(m)$	The i_{th} segment after dimension preprocessing
$c(k)$	Square spectra of $b(n)$
L_C	Input dimension of the Att-CNN
$\mathbf{x}$	Input signal waveform of the Att-CNN
$\mathbf{H}$	Filter group
$\mathbf{h}$	Single filter in $\mathbf{H}$
$\mathbf{U}$	The convolution output
$\mathbf{V}$	Channel-wise aggregated statistic
$\mathbf{F}_{sq}(\cdot)$	Squeezing function of SE unit
$\boldsymbol{\lambda}$	Excitation vector
L_F	Filter length
T	Temporal dimension of $\mathbf{x}$
C	Channel number of $\mathbf{H}$
$\mathbf{F}_{ex}(\cdot, \mathbf{W})$	Excitation function of the SE unit
$\mathbf{W}$	Weight of FC layers
r	Reduction ratio of FC layers
$\tilde{\mathbf{U}}$	Rescaled features
$\mathbf{F}_{scale}(\cdot, \cdot)$	Rescaling function
$\mathbf{h}_{filter}$	Complex filter constituted of multiple filter groups
$\tilde{\mathbf{x}}$	Network output with SE unit
$\hat{\mathbf{x}}$	Network output without SE unit
J	Total loss function during the pre-training of the SAE
$f_{SAE}(\cdot)$	Nonlinear function formed by the SAE
L_s	Number of input nodes of the SAE
L_n	Number of neurons of the SAE
ρ_i	Activation value of the i_{th} neuron
β_n	Sparsity penalty
ρ	Expected average activation value
P_d^i	The prediction probability vector of the i_{th} segment
P_d	Overall prediction probability vector of Att-CNN module
P_c	Prediction probability vector of the SAE module
P_v	Final prediction probability after weighting
λ_d	Weight of P_d
$\tilde{y}(n)$	Received signal in the transfer data model
$\tilde{h}(n)$	UWA channel in the transfer data model

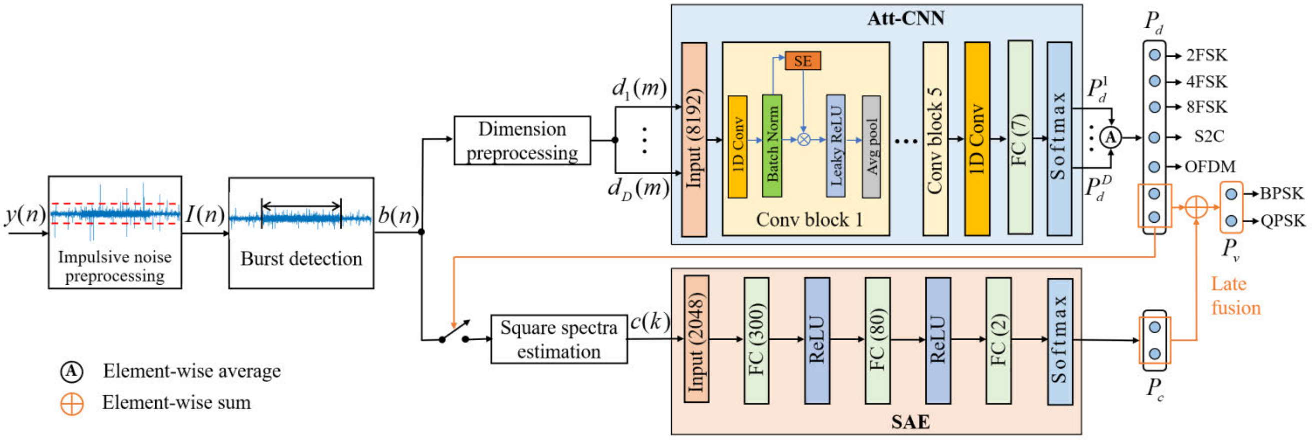


FIGURE 1. Architecture of the proposed AMC model based on HNNs.

TABLE 3. ***(Continued.)*** **List of the mathematical notions used in this paper.**

$\tilde{w}(n)$	Ambient noise in the transfer data model
$\tilde{\alpha}$	Characteristic exponent in the transfer data model
h_0	Impulsive response of channel $\tilde{h}(n)=1$
$H(z)$	Transfer function of $h(n)$
l_r	Learning rate
e_p	Training epochs
N	Number of transmitted symbols
N_T	Number of training examples per modulation

approach. Moreover, a transfer learning strategy is presented to overcome the problem of data scarcity at the end of this section.

A. SIGNAL MODEL

A UWA communication signal is affected by multi-path arrivals and marine ambient noise during its transmission. Thus, the received signal $y(n)$ can be modeled as follows:

$$y(n) = s(n) * h(n) + w(n), \tag{1}$$

where $y(n)$ has L samples, $s(n)$ represents the transmitted signal, with a modulation set including 2FSK, 4FSK, 8FSK, BPSK, QPSK, OFDM, and S2C, $*$ represents the convolution operator, $h(n)$ is the impulse response of the UWA channel, and $w(n)$ is the ambient noise.

Because of the frequent industrial and marine life activities, short, high-amplitude impulsive noise bursts are observed in shallow-water regions [30], [31]. Previous studies have shown that the distribution of this type of noise is closer to an alpha-stable distribution [31], [32]. Therefore, to better characterize the actual marine ambient noise, $w(n)$ is modeled as an alpha-stable distributed noise. Its characteristic function can be expressed as [33]:

$$\phi(u) = \exp\left(jau - \gamma\,|u|^{\alpha}\,[1 + j\beta sgn(u)\omega(u,\alpha)]\right), \tag{2}$$

where,

$$\omega(u,\alpha) = \begin{cases} \tan(\pi\alpha/2) & \alpha \neq 1 \\ (2/\pi)\lg|u| & \alpha = 1, \end{cases} \tag{3}$$

$$sgn(u) = \begin{cases} 1 & u > 0 \\ 0 & u = 0 \\ -1 & u < 0, \end{cases} \tag{4}$$

and $0 < \alpha \leq 2$, $-\infty < a < \infty$, $\gamma > 0$, and $-1 \leq \beta \leq 1$. The characteristic exponent α measures the intensity of the impulse, and the lower the value of α, the higher the intensity. The location parameter a represents the center axis of the distribution function. The dispersion γ is a measure of the distribution deviation around its mean value. The symmetry parameter β describes the skewness of the distribution function, and we have a symmetric alpha-stable ($S\alpha S$) distribution when $\beta = 0$. Furthermore, when $a = 0$ and $\gamma = 1$, $S\alpha S$ becomes a standard alpha-stable distribution.

Because an alpha-stable distribution has no second-order or higher-order statistics when $\alpha < 2$, the power relationship between the signal and noise can be measured using the mixed signal-to-noise ratio (MSNR), which can be expressed as:

$$\text{MSNR} = \left[10\lg\left(\sigma_s^2/\gamma\right)\right](\text{dB}), \tag{5}$$

where σ_s^2 denotes the variance of the signal.

B. PROPOSED AMC MODEL

Fig. 1 shows the architecture of the proposed AMC model based on HNNs. It mainly consists of an Att-CNN module, and an SAE module, in addition to the preprocessing part. First, an impulsive noise preprocessor (INP) is adopted to reduce the high-amplitude impulse in the received noisy signal $y(n)$. Second, burst detection is performed on $I(n)$, and the transmitted communication data block $b(n)$ is detected. Third, the dimension of $b(n)$ is preprocessed to match the input dimension of the proposed Att-CNN module. Fourth, the Att-CNN module performs preliminary classification on the temporal waveform of $d_i(m)$. If a signal is recognized as BPSK- or QPSK-modulated, the sequence of its estimated square spectra will be additionally fed to the proposed SAE module for inter-class recognition. Finally, a confidence-based late fusion is made based on the output classification probabilities of the two modules.

1) IMPULSIVE NOISE PREPROCESSING

Marine ambient noise has a wide dynamic range, particularly in the presence of the high-amplitude impulse. The significant numerical difference between different waveform samples increases the probabilities of gradient imbalance and model non-convergence during the network training. Thus, it is necessary to apply impulse reduction and normalization preprocessing on the received signals. The INP adopted in this study aims to nonlinearly suppress the locations where the amplitude is higher than the selected threshold τ_r in the received signal $y(n)$. The denoised signal can be represented as [34]:

$$y'(n) = \begin{cases} y(n), & |y(n)| \leq \tau_r \\ y(n)\left(\dfrac{\tau_r}{|y(n)|}\right)^2, & |y(n)| > \tau_r, \end{cases} \tag{6}$$

$$\tau_r = (1 + 2\tau_0)\tau_Q, \tag{7}$$

where τ_0 is a constant coefficient (e.g., $\tau_0 = 1.5$ is considered in this paper, and τ_Q is the second quartile of the absolute value of the received signal. Thereafter, $y'(n)$ is further normalized to obtain the final output of the INP:

$$I(n) = \frac{y'(n)}{\max(|y'(n)|)}. \tag{8}$$

2) BURST DETECTION

After impulsive noise preprocessing, the high-amplitude impulsive noise is significantly suppressed; however, there is still heavy low-amplitude noise in $I(n)$. The transmitted burst signal can be drown in noise, with its start and end time hard to detect. Thus, it is necessary to extract the useful communication data block from $I(n)$ for the subsequent classification. In this method, a trained signal denoising generative adversarial network (SDGAN) proposed in our previous work [35] is adopted to perform burst detection. SDGAN has a generator and a discriminator, which are trained adversarially to learn the target data distribution. During the training, the generator aims to generate data similar to the target data. The discriminator evaluates the quality of the generated data automatically using the metric learned itself, and offers a better updating direction to the generator. When $I(n)$ and $s(n)$ are taken as the source and target data, respectively, the trained generator will be able to suppress the noise component in $I(n)$ to nearly zero, with the communication data block well remained. Thus, the burst signal $b(n)$ in $I(n)$ can be well detected, and $b(n)$ has L_b samples ($L_b \leq L$).

3) DIMENSION PREPROCESSING

The proposed Att-CNN module has a fixed input dimension L_C, whereas the signal burst durations vary. Thus, the denoised temporal waveforms with a dimension L_b cannot be directly fed to the Att-CNN module. To handle the case when $L_b > L_C$, Zheng *et al.* [36] proposed a signal segmentation and fusion approach. After segmentation without overlapping, the i_{th} segment can be expressed as follows [36]:

$$d_i(m) = b((i-1)L_c + m), \tag{9}$$

where $m = 1, 2, \ldots, L_c$, $i = 1, 2, \ldots, D$, and $D = \lfloor L_b/L_C \rfloor$, $\lfloor \eta \rfloor$ denotes the largest integer that is not greater than η. Thereafter, all the D segments are fed to the classification network in turn, and multiple prediction probability vectors $P_d^1, \ldots, P_d^D$ can be obtained. Finally, these vectors are averaged, and a comprehensive judgement is made accordingly. This method is proven to help make full use of the information carried by long-duration signals and perform better. Thus, it is adopted in our study to perform dimension preprocessing when $L_b > L_C$.

Typically, if the computational cost is not a consideration, a larger L_C means more information can be utilized simultaneously to extract better temporal correlations for recognition. However, this increases the probability of $L_C > L_b$, particularly when short burst signals are received. To the best of our knowledge, few works proposed a solution to this problem. Therefore, in this paper, we present a data-reusing (DR) technique to extend the dimension L_b to L_C for short burst signals. DR is designed to repeatedly concatenate the sequence $b(n)$ in the time dimension. The output after DR is expressed as follows:

$$d_i(m) = \begin{cases} b(\mathrm{mod}(m/L_b)), & \mathrm{mod}(m/L_b) \neq 0 \\ b(L_b), & \text{else}, \end{cases} \tag{10}$$

where $i = 1, m = 1, 2, \ldots, L_c$, $\mathrm{mod}(\cdot)$ represents the modulo operator. In fact, $d_i(m)$ and $b(n)$ contain the same symbol and noise information, thus the modulation features remain approximately unchanged after DR.

4) ATTENTION AIDED CONVOLUTIONAL NEURAL NETWORK

A CNN has multiple filter banks with the characteristics of weight sharing and local receptive fields. Compared with other common neural networks, such as fully connected (FC) networks, it has fewer parameters and can better extract local features. In fact, the modulated signals are over-sampled from a string of independent symbols. Thus, there is a certain correlation between the samples within a single symbol. To better learn the temporally close correlations of the signal waveforms, an Att-CNN module is built mainly based on the structure of one-dimensional convolutional (1D Conv) layers. Moreover, the squeeze and excitation (SE) unit [37], which is an efficient self-attention mechanism, is introduced for performance improvement.

As shown in Fig. 1, considering the over-sampling rates of the different UWA communication signals used in the later experiments, the input dimension of the Att-CNN module L_C is set to 8192. This module first uses five convolutional blocks to extract the signal features from the input waveforms. Second, a 1D Conv layer with a single filter (filter length $L_F = 1$) is used to compress the feature channels, and a 1D feature vector can be obtained. Third, this feature vector is connected to an FC layer, which works as a classifier. Finally, the softmax activation function outputs the classification probability vector P_d of the seven modulations.

In each convolutional block, there is first a 1D Conv layer ($L_F = 31$) and a batch normalization layer for feature extraction and data normalization, respectively. The convolutional layers in the different blocks have 16, 32, 64, 128, and 256 filters. Thus, multiple channels of features can be obtained by these 1D Conv layers. Thereafter, an SE unit is introduced to learn a weight vector for the features of the different channels and recalibrates them. Finally, a leaky ReLU function and an average pooling layer with a stride of 4 (except convolutional block 5) are used for nonlinear activation and network compression, respectively.

The SE unit utilized in the convolutional blocks is a popular channel attention mechanism in the field of computer vision. In this study, it is transferred and used for the processing of 1D temporal waveforms. For the sake of analysis, we take a simplified version of the proposed Att-CNN module as an example. As shown in Fig. 2, the number of convolutional blocks is reduced to one, and only the main structures within a block (i.e., the 1D Conv layer and SE unit) are considered. Moreover, we assume that the two 1D Conv operations (i.e., Conv1 and Conv2) do not affect the time dimension (i.e., T), with their bias term omitted. Based on the operation mechanism of the SE unit in [37], a detailed derivation from the principal of signal filtering is given to prove its advantage for the signal feature extraction.

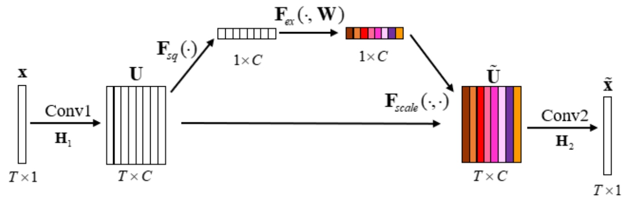


FIGURE 2. **Simplified version of the proposed Att-CNN module.**

As shown in Fig. 2, the input signal waveform $\mathbf{x} \in \mathbb{R}^{T\times 1}$ is first convoluted with a filter bank $\mathbf{H}_1 \in \mathbb{R}^{L_F\times C}$ composed of C filters $(\mathbf{h}_{11}, \mathbf{h}_{12}, \ldots, \mathbf{h}_{1C})$. The output $\mathbf{U} \in \mathbb{R}^{T\times C}$ can be expressed as:

$$\begin{aligned}\mathbf{U} &= \mathbf{x} * \mathbf{H}_1\\ &= \mathbf{x} * [\mathbf{h}_{11}, \mathbf{h}_{12}, \ldots, \mathbf{h}_{1C}]\\ &= [\mathbf{x} * \mathbf{h}_{11}, \mathbf{x} * \mathbf{h}_{12} \ldots \mathbf{x} * \mathbf{h}_{1C}].\\ &= [\mathbf{u}_1, \mathbf{u}_2 \ldots \mathbf{u}_C] \end{aligned} \tag{11}$$

Equation (11) illustrates that the different channels of the features are obtained through the independent convolution of multiple 1D filters in $\mathbf{H}_1$. To increase the sensitivity of more important features, it is necessary to weight these features with the global information obtained. The SE unit is designed to achieve this in two steps: squeezing and excitation.

Because each filter in $\mathbf{H}_1$ operates with a local receptive field whose length $L_F < T$, it cannot exploit the contextual information outside of this region. To this end, the global average pooling operation is adopted to squeeze the global temporal information in each channel. Thus, a channel-wise aggregated statistic $\mathbf{V} \in \mathbb{R}^C$ is obtained and can be expressed as:

$$\mathbf{V} = \mathbf{F}_{sq}(\mathbf{U}) = \frac{1}{T}[\sum_{i=1}^{T}\mathbf{u}_1(i), \ldots, \sum_{i=1}^{T}\mathbf{u}_2(i), \ldots, \sum_{i=1}^{T}\mathbf{u}_C(i)] \tag{12}$$

Based on this, we continue to seek for channel-wise dependencies and calculate the excitation vector $\lambda \in \mathbb{R}^C$ for different channels (i.e., the channel weight). This is achieved with two FC layers [37]:

$$\lambda = \mathbf{F}_{ex}(\mathbf{V}, \mathbf{W}) = \sigma(\mathbf{W}_2\delta(\mathbf{W}_1\mathbf{V})), \tag{13}$$

where σ and δ denote the Sigmoid and ReLU functions, respectively, $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r}\times C}$ and $\mathbf{W}_2 \in R^{C\times\frac{C}{r}}$ are the weights of the two FC layers used for decreasing and increasing the dimension, respectively. The reduction ratio r is adapted to keep the middle layer dimension $\frac{C}{r}$ at 8 in our study. Thereafter, the learned excitation vector $\lambda \in R^C$ is utilized to weight the different filter responses, and a rescaled $\tilde{\mathbf{U}} \in \mathbb{R}^{T\times C}$ is obtained, which is expressed as:

$$\tilde{\mathbf{U}} = \mathbf{F}_{scale}(\mathbf{U}, \lambda) = [\lambda_1\mathbf{u}_1, \lambda_2\mathbf{u}_2, \ldots, \lambda_C\mathbf{u}_C] \tag{14}$$

Therefore, the features of the different channels can be emphasized or suppressed on the basis of their correlations with the final objectives. Finally, $\tilde{\mathbf{U}}$ is convoluted with a single 2D filter $\mathbf{H}_2 \in \mathbb{R}^{C\times L_F}$. The output can be expressed as:

$$\begin{aligned}\tilde{\mathbf{x}} &= \tilde{\mathbf{U}} * \mathbf{H}_2 = [\lambda_1\mathbf{u}_1, \lambda_2\mathbf{u}_2, \ldots, \lambda_C\mathbf{u}_C] * \begin{bmatrix}\mathbf{h}_{21}\\ \mathbf{h}_{22}\\ \cdots\\ \mathbf{h}_{2C}\end{bmatrix}\\ &= \lambda_1\mathbf{u}_1 * \mathbf{h}_{21} + \lambda_2\mathbf{u}_2 * \mathbf{h}_{22} + \ldots + \lambda_C\mathbf{u}_C * \mathbf{h}_{2C}\\ &= \lambda_1\mathbf{x} * \mathbf{h}_{11} * \mathbf{h}_{21} + \lambda_2\mathbf{x} * \mathbf{h}_{12} * \mathbf{h}_{22}\\ &\quad + \ldots + \lambda_C\mathbf{x} * \mathbf{h}_{1C} * \mathbf{h}_{2C}\\ &= \mathbf{x} * (\lambda_1\mathbf{h}_{11} * \mathbf{h}_{21} + \lambda_2\mathbf{h}_{12} * \mathbf{h}_{22} + \ldots + \lambda_C\mathbf{h}_{1C} * \mathbf{h}_{2C})\\ &= \mathbf{x} * \mathbf{h}_{filter} \end{aligned} \tag{15}$$

Based on the above analysis, it can be concluded that the multiple filter banks of the different layers constitute a more complex filter $\mathbf{h}_{filter}$. Moreover, owing to the SE unit and the learned λ, this filter can be adaptively updated as the input varies. Therefore, different influence functions can be built for feature extraction on different testing data. In comparison, the network output without the SE unit is expressed as:

$$\hat{\mathbf{x}} = \mathbf{x} * (\mathbf{h}_{11} * \mathbf{h}_{21} + \mathbf{h}_{12} * \mathbf{h}_{22} + \ldots + \mathbf{h}_{1C} * \mathbf{h}_{2C}). \tag{16}$$

The network has a fixed influence function for the different input data. In fact, the frequency spectrum characteristics vary with the signals, and different filters can effectively extract the different frequency-domain features, owing to their different frequency responses. Thus, the SE unit has evident advantages for signal feature extraction.

The above conclusion also applies to the proposed Att-CNN module. After it performs the preliminary recognition, the prediction probability vector P_d of the seven modulations is obtained. If $L_b > L_C$, based on the signal segmentation and fusion method [36] introduced in the dimension preprocessing subsection, the multiple prediction probability vectors of different segments $(P_d^1, \ldots, P_d^D)$ can be averaged to obtain a final P_d. If the maximum probability value of P_d corresponds to 2FSK, 4FSK, 8FSK, S2C, or OFDM, the judgement is directly made. Otherwise, the SAE module, shown in Fig. 1, will be further adopted for PSK inter-class recognition. This can be attributed to the vulnerability of the phase information carried by the temporal waveforms, which is not robust enough under complex marine environments.

5) SAE FOR PSK INTER-CLASS RECOGNITION

The square spectra of the BPSK signal has an evident impulse at the double-carrier frequency, whereas that of the QPSK signal does not. Fig. 3 shows the examples of their square spectra (the carrier frequency is 12.5 kHz). Hence, this characteristic difference is used to distinguish between them in our study. The Welch method is adopted to calculate the square spectra of $I(n)$ with a window length of 2048. Thereafter, a sequence $c(k), k = 1, 2, \ldots, 2048$ of the estimated square spectra is obtained and fed to the SAE.

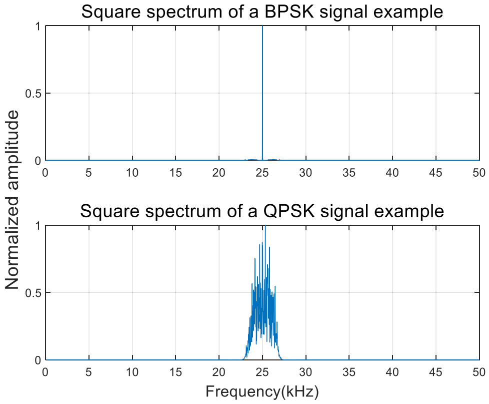


FIGURE 3. Square spectra of BPSK and QPSK signal examples.

As shown in Fig. 1, the SAE module has 2048 input nodes, and the two hidden layers have 300 and 80 nodes, respectively. The final FC layer is connected to a softmax function, which outputs the binary classification probability vector P_c. The ReLU function is adopted for activation in the SAE module. The training of the presented SAE module involves two steps: unsupervised pre-training and supervised training. The former enables the network to learn some low-dimensional features by training it to minimize the reconstruction error of the input data. Meanwhile, a sparsity constraint is applied by setting the average activation values of the neurons. Thus, the SAE module is able to learn more sparse features. The total loss function during the pre-training stage can be expressed as [38]:

$$J = \frac{1}{L_s}\sum_{k=1}^{L_s}(\frac{1}{2}\|f_{SAE}(c(k)) - c(k)\|_2^2) + \beta_n \sum_{i=1}^{L_n} KL(\rho||\rho_i), \tag{17}$$

$$KL(\rho||\rho_i) = \rho \log\frac{\rho}{\rho_i} + (1-\rho)\log\frac{1-\rho}{1-\rho_i}, \tag{18}$$

where $f_{SAE}(\cdot)$ is the nonlinear function formed by the SAE network, L_s is the number of input nodes (i.e., 2048), L_n is the number of neuros, ρ_i is the activation value of the i_{th} neuron, and the weight of the sparsity penalty β_n and the expected average activation value ρ are set to 3 and 0.05, respectively.

The supervised training step of the SAE module is similar to that of the Att-CNN module. The errors between the predicted values and the preset labels are calculated and the network parameters are updated through the back-propagation algorithm. Finally, effective classification features can be learned when the training is completed.

6) LATE FUSION

The two aforementioned modules are designed to perform multi-class and binary classification, respectively. Thus, they are trained individually. When it comes to network testing, if a signal is recognized as BPSK- or QPSK-modulated, a fusion is further required to fuse the two modules. Common fusion approaches are mainly divided into three categories: early, late, and hybrid fusion [24], [25]. However, only late fusion is available and considered in this study. There are three reasons: (i) The SAE module is only used when a signal is judged as PSK-modulated by the Att-CNN module, i.e., the case that one modality is missing exists, whilst early and hybrid fusion methods require all the modalities; (ii) The two modules have different classification categories, whilst the module tasks are required to be the same for the other two fusion methods; (iii) The length of $b(n)$ is variable, i.e., the Att-CNN module is implemented for unknown times on different signal segments, whilst the other two fusion methods require parallel data for the two modalities. These factors made the early fusion approaches, such as feature concatenation [26], [27], and hybrid fusion approaches [28] unavailable. Finally, a confidence-based late fusion is further adopted to fuse the output of the two modules, and an overall prediction probability vector for PSK signals can be expressed as:

$$P_{v-PSK} = \lambda_d P_{d-PSK} + (1-\lambda_d)P_{c-PSK}, \quad \lambda_d \in [0, 1], \tag{19}$$

where P_{d-PSK} and P_{c-PSK} are the probability vectors of the PSK signals predicted by the Att-CNN and SAE module, respectively, and λ_d and $(1-\lambda_d)$ are their corresponding weights. The later experiments indicate that a too high or low value of λ_d will deteriorate the performance at low or high SNRs, respectively. Overall, a relatively best performance is obtained when $\lambda_d = 0.5$.

7) TRANSFER LEARNING

The square spectra of PSK signals represent the statistical characteristics of the signal phase. By contrast, the temporal waveforms show the instantaneous information of the signals, and thus are more vulnerable to UWA channels. Therefore, for the proposed Att-CNN module, the training data distribution is expected to be the same to that of the testing data. However, because of the sparsity of UWA communication signals, it is difficult to acquire enough training data from a testing channel to train a reliable network. To this end, we introduce the idea of transfer learning and present a transfer data model:

$$\tilde{y}(n) = s(n) * \tilde{h}(n) + \tilde{w}(n), \tag{20}$$

where $\tilde{y}(n)$ is the received signal, $s(n)$ is the transmitted signal with the same modulation set as in (1), $\tilde{w}(n)$ is also modeled as an alpha-stable distributed noise with the characteristic exponent $\tilde{\alpha}$, and $\tilde{h}(n)$ denotes a channel similar to the testing channel, such as a channel obtained under different transmitting depths or distances in the same water region. When a similar channel is unavailable, $\tilde{h}(n) = 1$ is generally taken and denoted as channel h_0.

Although the data distributions in (1) and (20) are different, $y(n)$ and $\tilde{y}(n)$ contain the same components, i.e., the transmitted signal waveforms. Thus, the transfer is reasonable and feasible. Moreover, a two-step training method is adopted to conduct the transfer learning strategy in our study. Fig. 4 shows the overall training process.

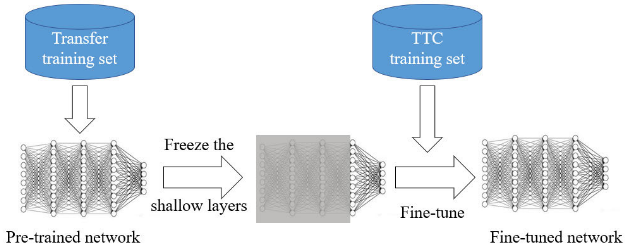


FIGURE 4. Two-step training process of the transfer learning strategy.

As shown in Fig. 4, the transfer learning strategy is conducted in two steps: pre-training on a transfer training set and fine-tuning on a target testing channel (TTC) training set. The transfer training set is based on the transfer data model in (20) and contains large amounts of data. In comparison, the TTC training set has a limited amount of data from the testing channel. However, because of the data scarcity, the latter step is likely to cause over-fitting. A common solution is to only fine-tune the parameters of the last few layers, while freezing those of the shallow ones. It is demonstrated in [39] that the features learned by the shallow layers of a network can be generalized. However, the deeper layers can extract more specific features, which are more targeted to the input data. Consequently, all the parameters of the Att-CNN module are frozen except for those of the last few layers during the fine-tuning stage in our study.

III. NUMERICAL RESULTS AND DISCUSSION

A. SIGNAL PARAMETERS AND DATASET

All signal examples are generated on the basis of the two aforementioned data models in (1) and (20), with a sampling rate of 48 kHz. Their carrier frequencies randomly vary in the range of 15 to 16 kHz, except for those of S2C in [8, 12] kHz. The sub-carriers of the OFDM signals are randomly modulated with BPSK or QPSK, and the PSK signals are shaped using root-raised-cosine pulse-shaping filters. Table 4 lists the other parameters.

Moreover, the widely used Bellhop channel simulation software is adopted to generate different UWA channels based on the popular Argo ocean database. The hydrologic data at the coordinates (165.5°E,45.5°N) are used to generate six sparse channels under different transmission conditions. Fig. 5 shows the sound velocity profile of this water region. Table 5 lists the detailed channel parameters.

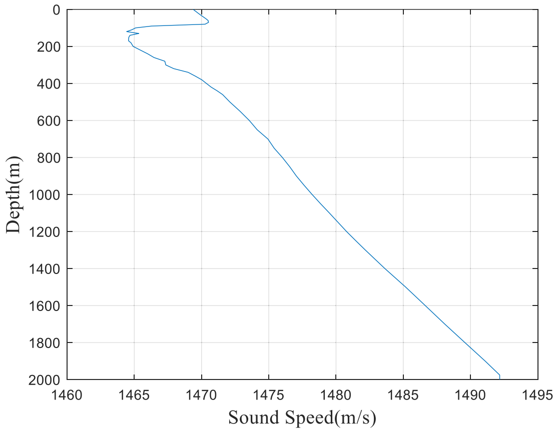


FIGURE 5. Sound velocity profile.

The transfer functions of the above six channels are:

$$\begin{aligned}
H_A(z) &= 0.04 + z^{-353} + 0.508z^{-570} + 0.283z^{-644},\\
H_B(z) &= 0.32 + 0.45z^{-48} + z^{-61} + 0.9318z^{-267},\\
H_C(z) &= 0.68 + z^{-184} + 0.882z^{-403},\\
H_D(z) &= 0.177 + 0.265z^{-534} + z^{-689} + 0.369z^{-800},\\
H_E(z) &= 0.606 + 0.49z^{-741} + 0.878z^{-2613} + z^{-3535},\\
H_F(z) &= 0.5577 + 0.4213z^{-922} + z^{-4621} + 0.8755z^{-5739}.
\end{aligned}$$

Because the sampling rate is set to 48 kHz, the maximum propagation delays of these channels are 13.4, 5.5, 8.4, 16.7, 73.6, and 119.6 ms, respectively. Furthermore, Fig. 6 shows the amplitude–frequency response curves of these channels.

TABLE 4. Signal parameters.

Signal types	Symbol rate (Baud)	Modulation index	Roll-off factor	Cyclic prefix	No. Subcarriers	Sweeping bandwidth (Hz)
2FSK	[530, 1k]	1	/	/	/	/
4FSK	[320, 600]	1	/	/	/	/
8FSK	[170, 330]	1	/	/	/	/
OFDM	{150, 200, 240, 320}	/	{0.2, 0.25, 0.3}	0.25	{4, 8, 16}	/
MPSK	{1.6k, 2k, 2.4k, 3k}	/	{0.2, 0.25, 0.3}	/	/	/
S2C	[400, 1k]	/	/	/	/	[8k, 12k]

TABLE 5. Parameters of different UWA channels.

Channels	Transmitter depths (m)	Distances (km)	Receiver depths (m)
h_A	200	5	200
h_B	100	3	200
h_C	200	5	50
h_D	200	8	200
h_E	50	15	120
h_F	50	12	200

As shown in Fig. 6, the frequency-selective fading characteristics of the above channels are different, and the declines in the channel h_E and h_F are relatively deeper. Moreover, different channels are used to generate different training and testing datasets in subsequent experiments. In each training set, if not specified, 6000 examples are generated for each modulation, with $\tilde{\alpha}$ and MSNR in the ranges of [1.5, 2] and [–10, 20] dB, respectively. For testing data, the communication data block occupies half the duration of the processing data block (i.e., the received signal).

The network training and testing are conducted with the DL library, PyTorch, on a single NVIDIA TITAN RTX GPU. The Adam optimizer [40] is used to optimize the network parameters until the loss function converges with a batch size of 128. The learning rate l_r gradually decreases from an initial value $l_0 = 0.01$, and l_r can be expressed as:

$$l_r(e_p) = \begin{cases} 0.5^{\lfloor e_p/5 \rfloor} l_0, & 0.5^{\lfloor e_p/5 \rfloor} l_0 > 10^{-5} \\ 10^{-5}, & else, \end{cases} \tag{21}$$

where, e_p denotes the training epochs.

B. SIMULATION EXPERIMENTS AND DISCUSSION

1) PERFORMANCE COMPARISON

First, to prove the effectiveness of the proposed method for short burst UWA communication signals classification, its performance is compared with those of its variants. The training and testing datasets are generated under channel h_A and $\alpha = 1.5$ is used for testing. The MSNR of the testing examples varies from −6 to 16 dB with an interval of 2 dB. and 400 testing examples are generated for each modulation under different MSNRs. Moreover, there are 64 symbols within a single testing example. Fig. 7 shows the comparison results.

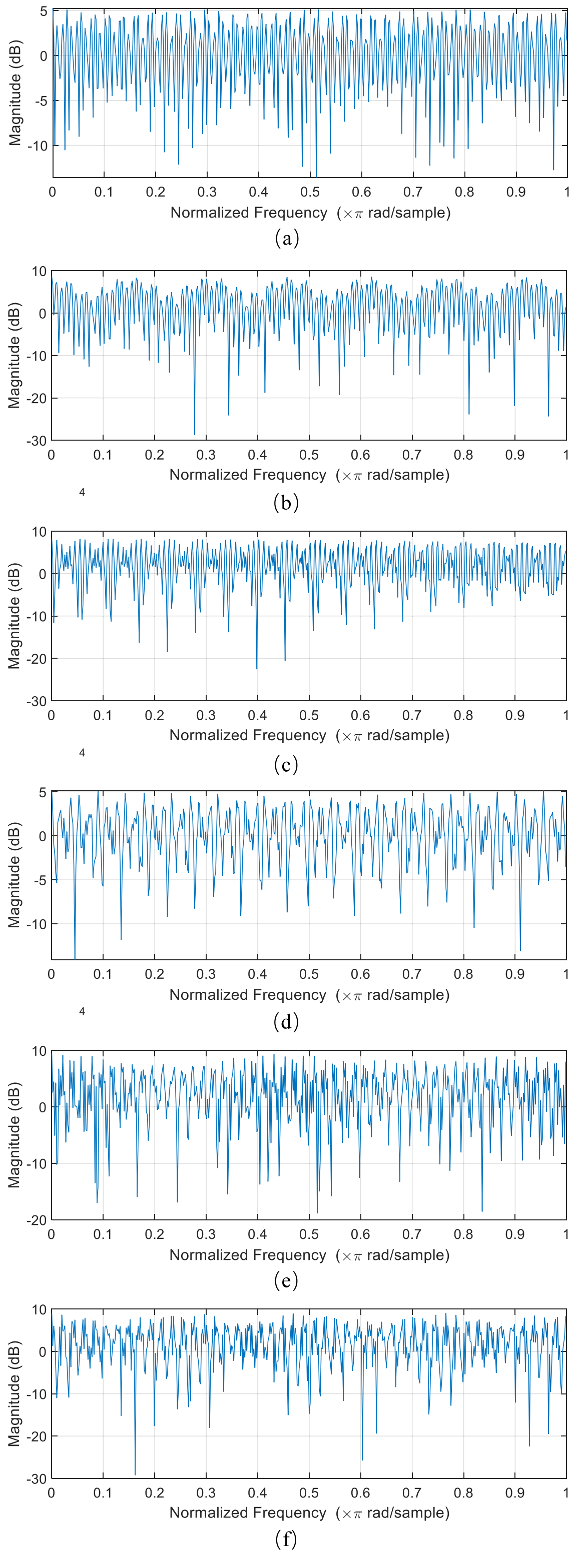
The CNN in Fig. 7 represents the recognition method using a common CNN network without attention. Moreover, instead of the DR method, it adopts a zero-padding (ZP) technique for dimension preprocessing. ZP pads a total of $L_c - L$ zero samples randomly on both sides of $b(n)$.The output after ZP can be expressed as follows:

$$d_i(m) = \begin{cases} b(m - m_b + 1), & m = m_b, m_b + 1, \ldots, m_b + L - 1 \\ 0, & \text{else}, \end{cases} \tag{22}$$

where $i = 1, m = 1, 2, \ldots, L_c$, m_b is a random integer in the range of $[1, L_c - L_b + 1]$. Moreover, the CNN–DR method replaces ZP with DR, and the Att-CNN–DR method (i.e. the proposed Att-CNN module) further adds the SE unit to CNN–DR. Att-CNN–DR–SAE is the proposed method using HNN.

As shown in Fig. 7, compared with the CNN method, the accuracy of the CNN–DR method decreases gradually below an MSNR of –2 dB, whereas it increases significantly at high MSNRs. In fact, the DR operation does not change the SNR of the signals essentially. However, the ZP operation can be approximately explained as padding data with an SNR of 0 dB, since the signal and noise powers of the padded samples are both zero. Therefore, the DR operation has a better performance than ZP at high MSNRs, whereas it deteriorates at low MSNRs. Fortunately, the adoption of the SE unit can effectively compensate for the insufficient performance of the CNN–DR method at low MSNRs. Actually, the CNN and Att-CNN have 1.35M and 1.36M trainable parameters, respectively. The slight increase in parameters has brought significant improvement in performance, which proves the high efficiency of the SE unit.

However, because the phase information represented in the temporal waveforms are not robust enough against UWA channels, the Att-CNN–DR method cannot effectively differentiate between BPSK and QPSK signals. Thus, this approach still encounters a performance bottleneck even at high MSNRs. When the SAE module is further adopted to extract more robust spectral features from the square spectra,



(a)

(b)

(c)

(d)

(e)

(f)

FIGURE 6. **Amplitude–frequency response curves of different UWA channels: (a) h_A; (b) h_B; (c) h_C; (d) h_D; (e) h_E; (f) h_F.**

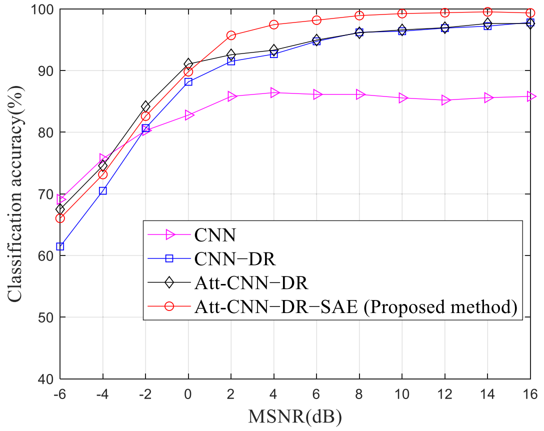


FIGURE 7. **Classification performances of the proposed method and its variants.**

the performance continues to improve at high MSNRs. The classification accuracy reaches approximately 99% at an MSNR of 8 dB. However, because the square operation has amplified the noise, the quality of the estimated square spectra declines quickly as MSNR decreases. Therefore, the performance of the proposed method is slightly worse than that of the Att-CNN–DR method below an MSNR of 0 dB. The number of trainable parameters increased to ∼2.0M, which is still acceptable.

Second, to prove the superiority of the proposed method, its performance is compared with those of the SAE-based method (SAE-2048) in [21] and the CNN-based method (CNN-1024) in [16]. The SAE-2048 approach adopts the Welch method for spectra estimation, and the window length is set to 2048. The two SAE networks adopted in this method each have two hidden layers, with 800 and 200 nodes, respectively. Moreover, all the seven modulations in (1) are included in the training and testing datasets. Each testing example contains 64 symbols. However, in the comparison with the CNN-1024 method, 4FSK, 8FSK, and OFDM are excluded. These signals have high over-sampling rates, whereas the CNN adopted in the CNN-1024 approach has a low input dimension of 1024 × 2. Thus, the corresponding number of symbols is too small to support the classification of the three modulations. The CNN in the CNN-1024 approach comprises two convolutional layers and two FC layers. The first convolutional layer has 64 filters of size 3 × 1, and the second one has 16 filters of size 3 × 2. The two FC layers have 128 and 4 nodes, respectively. We fed this CNN with the real and imaginary parts of signals. Moreover, each testing example contains 1024 samples, and the dimension is further extended to 8192 in our method. In the comparison with the above two methods, the training and testing datasets are built under channel h_A and $\alpha = 1.5$ is used for testing. Fig. 8 shows the results.

As shown in Fig. 8, the proposed method significantly outperforms the two compared methods, which have performance bottlenecks at high MSNRs. The classification accuracies are improved by 23% and 17%, respectively, compared with the SAE-2048 and CNN-1024 method at an MSNR of 10 dB. In fact, because of the insufficient symbols and fading characteristic of the channel, the quality of the estimated power spectra cannot be ensured for the

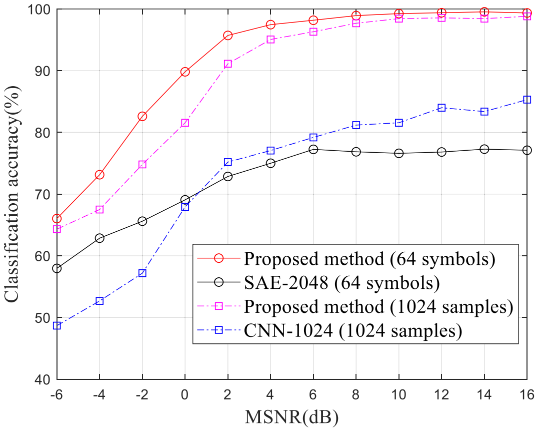


FIGURE 8. Performance comparison of different methods.

SAE-2048 method. Moreover, the structure of the CNN in the CNN-1024 method is too simple to recognize signals under complex environments. By contrast, the proposed method is able to make full use of the information carried by the limited waveform samples. Thus, it performs more robustly on short burst communication signals under UWA channels.

Table 6 lists the number of trainable parameters and the computational cost of the proposed HNN model and the considered DL baselines. The proposed HNN model has ∼2.0M trainable parameters and requires ∼0.1415 GFLOPs in a single forward pass. Compared with the SAE-2048 and CNN-1024 methods, it has fewer parameters; however, the computational burden is increased. The SAE-2048 method requires two additional spectral estimation operations, whereas only one is required in our method. Moreover, compared with the low input dimension of the CNN-1024 method, our method has a relatively high input dimension of 8192. Although this has increased the computational cost, more data can be simultaneously used to support the classification, and the computational cost is still acceptable.

TABLE 6. Model parameters and computational cost.

Models	Trainable parameters	GFLOPs
Proposed HNN	2.00 M	0.1415
SAE-2048	3.60 M	0.0072
CNN-1024	2.10 M	0.0122

2) CLASSIFICATION PERFORMANCE UNDER DIFFERENT CONDITIONS

In this section, we evaluate several factors that may affect the performance of the proposed method, including the burst duration, the intensity of the impulsive noise, the number pf training examples, and the prediction probability weight λ_d.

First, to evaluate the influence of burst duration, the classification performance is tested on signals containing different numbers of symbols under channel h_A. Four testing datasets are built, with the number of transmitted symbols $N = 32$, 64, 128, and 256, and α is set to 1.5. Fig. 9 shows the results.

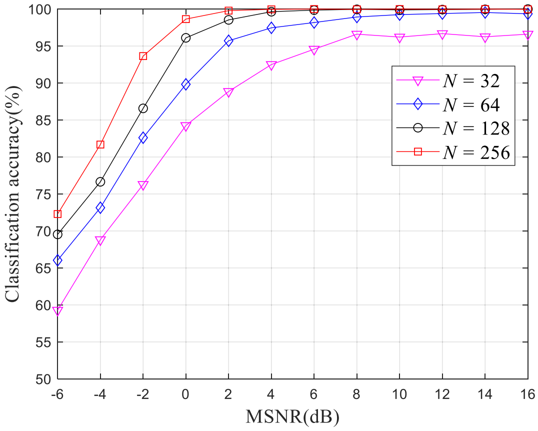


FIGURE 9. Classification performance when using different numbers of testing symbols.

As shown in Fig. 9, the classification accuracy increases quickly with the growing number of available symbols. In fact, the increase in the symbols helps the filter groups in the proposed Att-CNN module to extract better temporally correlation features, as indicated in (15). The quality of the estimated square spectra is also improved. Both these factors result in a better classification performance.

Second, to evaluate the influence of impulse intensity, the performance of the proposed method is tested on datasets with different α values, including 0.6, 0.9, 1.2, 1.5, and 1.8. These testing datasets are generated under channel h_A, and each testing example contains 64 symbols. Fig. 10 shows the results.

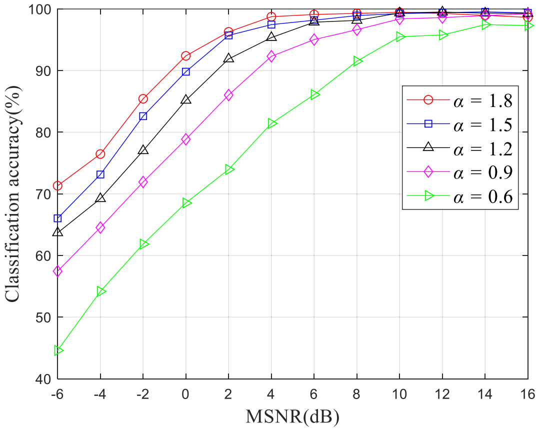


FIGURE 10. Classification performance under alpha-stable distributed noise with different impulse intensities.

As shown in Fig. 10, with the decrease in α(i.e., the increase in impulse intensity), the classification performance declines quickly. When the value of α decreases from 1.8 to 0.9, the accuracy decreases by approximately 13.5% at an MSNR of 0 dB. Nevertheless, benefited by the INP, the accuracy is still close to 80%. In fact, α is estimated to be in the range of [1.6, 1.8] for most actually observed noise examples in different water regions [41], [42]. Thus, the result indicates that the proposed method is robust against impulsive noise environment in most cases.

Third, to evaluate how the number of training examples affects the classification performance, different training datasets are generated under channel h_A. The number of training examples per modulation N_T is set to 500, 1000, 2000, 4000, and 6000, respectively. Moreover, the testing α value is set to 1.5, and each testing example contains 64 symbols. Fig. 11 shows the results.

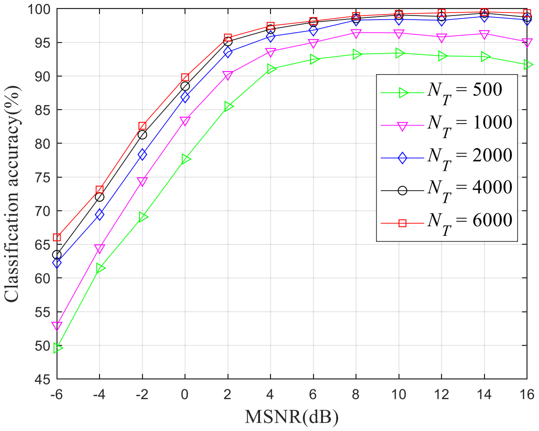


FIGURE 11. Classification performance when different numbers of training data are utilized.

As shown in Fig. 11, the classification accuracy increases with the growing number of training data, though with a decreasing rate. When N_T approaches 6000, the classification accuracy gradually stabilizes. This indicates that such a large amount of data is enough to support the network to fully learn the signal features. Thus, $N_T = 6000$ is adopted in this study to build the training dataset.

Finally, to evaluate the impact of the prediction probability weight λ_d, a performance comparison is made under channel h_A. The testing α value is set to 1.5, and each testing example contains 64 symbols. Fig. 12 shows the results when different values of λ_d are used.

As represented in (19), with the increase in λ_d, the prediction made by the Att-CNN module contributes more to the final decision. This leads to a better performance at low MSNRs, whereas a poorer performance at high MSNRs, as shown in Fig. 12. This result is consistent with the comparison result between the Att-CNN–DR and Att-CNN–DR–SAE methods shown in Fig. 7. Overall, the performance is relatively optimum when λ_d is set to 0.5.

3) TRANSFER LEARNING PERFORMANCE

In this section, the effectiveness of the proposed transfer learning strategy is demonstrated. First, experiments are conducted to prove the necessity of pre-training when data from the testing channel are limited and the feasibility of pre-training with data from similar available channels. We build two transfer training sets based on the data from channel h_0, and the data from channels h_A, h_B, and h_C, respectively. The testing channels are h_E and h_F. For each, 50 labeled examples for each modulation are generated to build the TTC training set. The TTC training sets are further utilized to fine-tune the pre-trained models. The fine-tuned models are also compared with the models directly trained with the transfer training sets without pre-training. Fig. 13 shows the results.

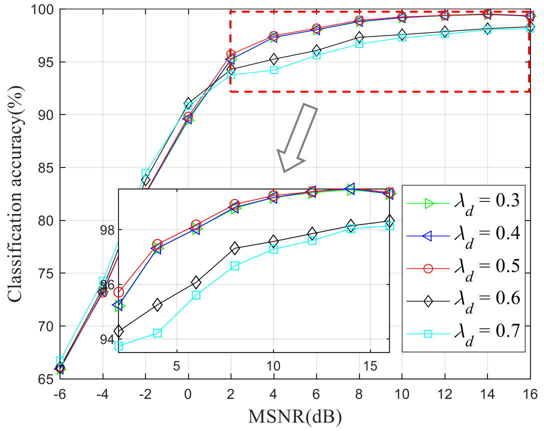


FIGURE 12. Classification performance under different λ_d.

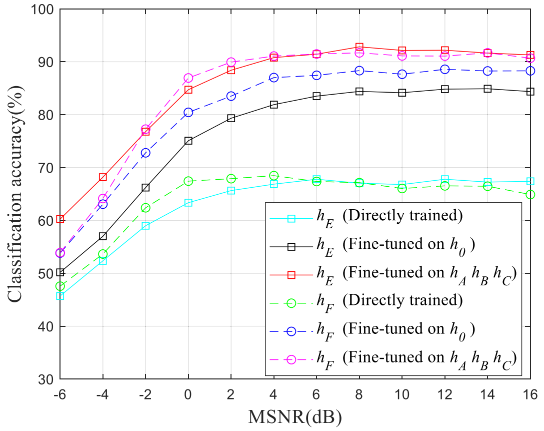


FIGURE 13. Classification performances of fine-tuned and directly trained models.

As shown in Fig. 13, for both the testing channel h_E and h_F, the performance of the directly trained model is worse than those of the fine-tuned models. The performance improves when data from similar channels (i.e., h_A, h_B, and h_C) are used for pre-training rather than an unrelated channel (i.e., h_0). The results show that the data from h_0, which does not carry any channel information, can help the network learn some general information of the signals themselves. However, using data from similar channels for pre-training can further help the network to learn some information related to the testing channel. This is key to the performance improvement. Thus,

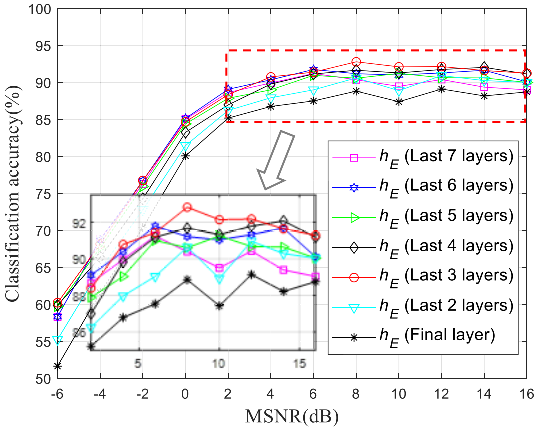


FIGURE 14. Classification performance of models obtained by fine-tuning different layers.

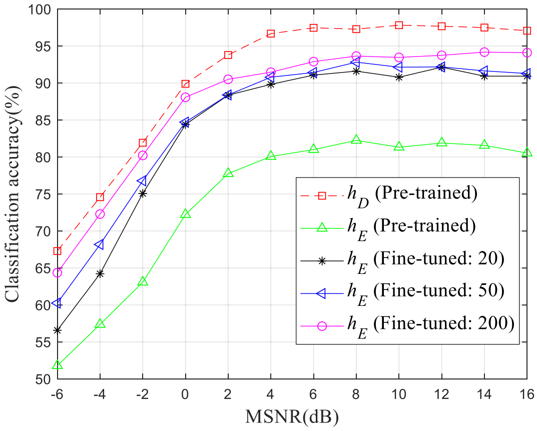


FIGURE 15. Classification performances of the pre-trained and fine-tuned networks.

the proposed transfer learning strategy is proven to be necessary and effective.

Second, we continue to evaluate the impact of fine-tuning different network layers on the transfer learning performance. The transfer training set is built on data from channels h_A, h_B, and h_C. And h_E is selected to be the testing channel. The TTC training set contains 50 examples for each modulation and is used to fine-tune the different layers of the Att-CNN module. Fig. 14 shows the results.

As shown in Fig. 14, the best performance is obtained when the last three layers are fine-tuned, and the performance decreases when a larger or smaller number of layers are fine-tuned. In fact, the features related to the channel h_E cannot be fully learned when fewer parameters are fine-tuned. Fine-tuning a large number of parameters with insufficient data will increase the probability of over-fitting. Therefore, the technique of fine-tuning the last three layers is proven to perform the best and is adopted in the other experiments when transfer learning is involved.

Finally, the influence of the amount of fine-tuning data on the transfer learning performance is evaluated. The same pre-trained model as in the last experiment is used to recognize the testing signals from the channels h_D and h_E. Moreover, we build several TTC training sets with different amounts of data from the channel h_E, including 20, 50, and 200 examples per modulation. Fig. 15 shows the classification performances of the pretrained and fine-tuned networks.

As shown in Fig. 15, the pre-trained network achieves a good performance under channel h_D, whereas it performs poorly under channel h_E. In fact, h_D has a shorter transmitting distance and more similar characteristics to the three training channels. However, h_E has significant differences, and the fading characteristics are poorer. Nevertheless, the performance under h_E is significantly improved even when fine-tuning with only 20 examples (a total duration of 3.4 s) per modulation. With the increase in the amount of data used for fine-tuning, the performance is gradually improved, though with a decreasing rate. The classification accuracy eventually stabilized at approximately 94%, which is lower than that under h_D(i.e., 97%). This can be attributed to the poor frequency response characteristics of h_E. Overall, the above experiments have demonstrated the effectiveness of the proposed transfer learning strategy and transfer data model.

C. PRACTICAL SIGNAL TESTS

To prove the effectiveness of the proposed method under actual marine environments, it is further tested on practical UWA communication signals. A lake trial and a sea trail are conducted in a lake on campus and in the Wuyuan Bay, Xiamen, respectively. Table 7 lists the channel parameters of the two tested water regions. Fig. 16 shows the experimental setup.

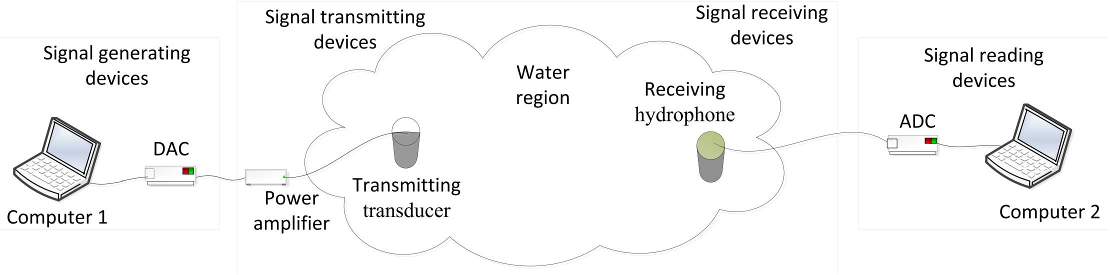
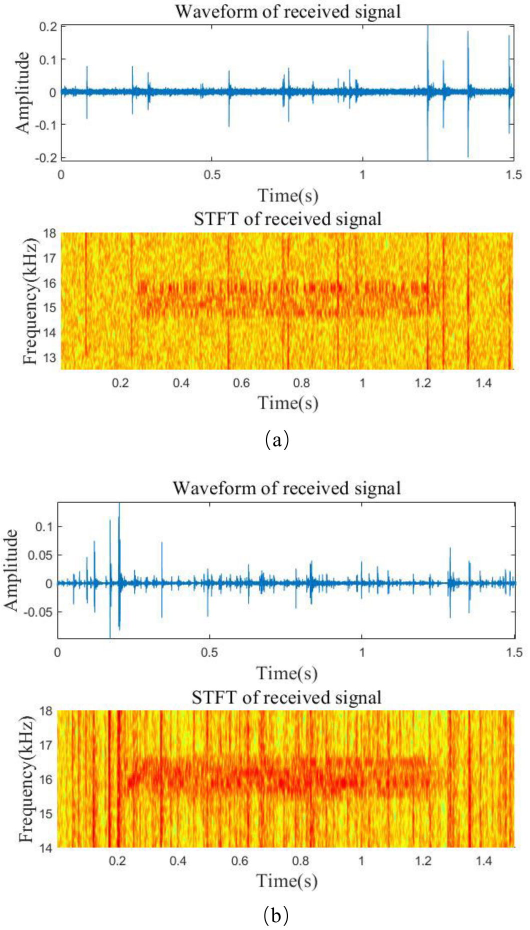


FIGURE 16. Experimental setup.



(a)

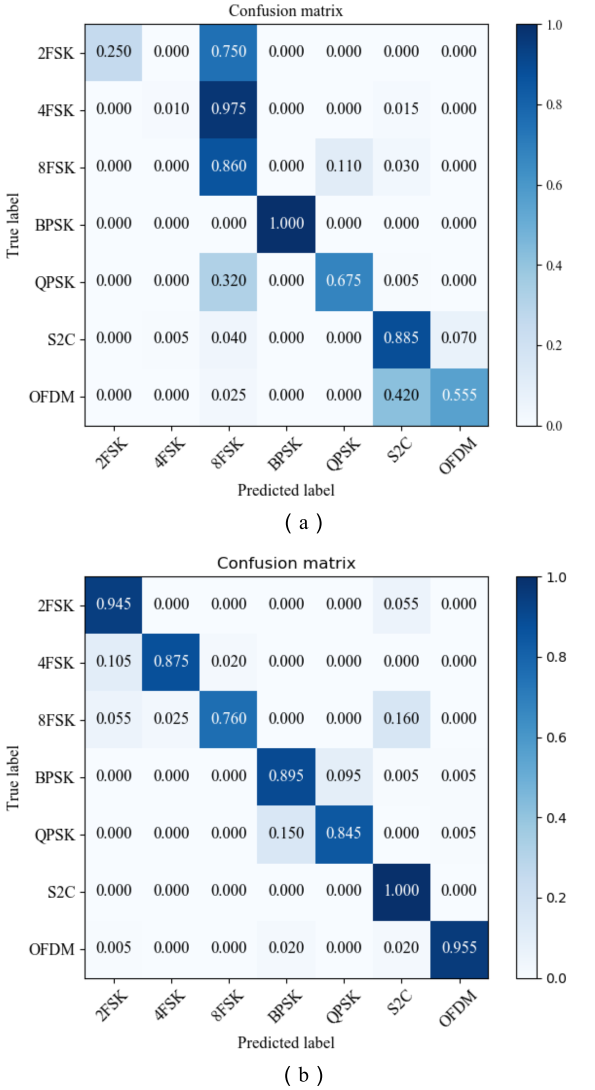
(b)

FIGURE 17. Temporal waveforms and STFT of 4FSK signal examples collected from: (a) Artificial lake; (b) Wuyuan Bay.

TABLE 7. Channel parameters of different water regions.

Water regions	Transmitter depth (m)	Receiver depth (m)	Distance (m)	Water depth (m)
Artificial lake	0.8	0.8	50	1-3
Wuyuan Bay	3	3	545	6-8

As shown in Fig. 16, the experiment setup is mainly divided into two parts: a transmitting part and a receiving part. The transmitting part consists of signal generating and transmitting devices. The receiving part consists of signal receiving and reading devices. The transmitting node is a common omnidirectional transducer and the receiving node is a single broadband hydrophone. The utilized hydrophone is the RB9-ETH model (Ocean Sonics). The sampling rate is set to 64 kHz, and the corresponding receiving bandwidth ranges from 10 Hz to 25.6 kHz. However, for trained networks, the testing data should be in the same format as the training data, i.e., their sampling rates should match. Thus, the received signals were resampled before testing.



(a)

(b)

FIGURE 18. Confusion matrices of the different methods in the artificial lake experiment. Note that the vertical and horizontal coordinates represent the true and predicted labels, respectively. (a) SAE-2048 method: The classification accuracies are low for most signals, e.g., 75% 2FSK, 97.5% 4FSK and 32% QPSK signal examples are misjudged as 8FSK-modulated, and 42% OFDM signal examples are misjudged as S2C-modulated, while the classification accuracies of 8FSK, BPSK and S2C are over 86%; (b) Proposed method: The classification accuracies reach over 84.5% for most signals, except when 8FSK has a relatively low accuracy of 76%.

During each experiment, the SNRs of the received signals were kept at a low level by reducing the transmitting power. Taking 4FSK modulation as an example, we present the temporal waveforms and short-time Fourier transformation (STFT) of two signal examples collected from the artificial lake and Wuyuan Bay, respectively, as shown in Fig. 17.

As shown in Fig. 17, there are evident intense impulsive noise in the received signals, and the SNR is low. The characteristic exponent α is estimated to be in the range of [1.54, 2.0] for the signals collected from the artificial lake, with the method of sample fractiles proposed in [43]. In the Wuyuan Bay experiment, α is estimated to be in the range

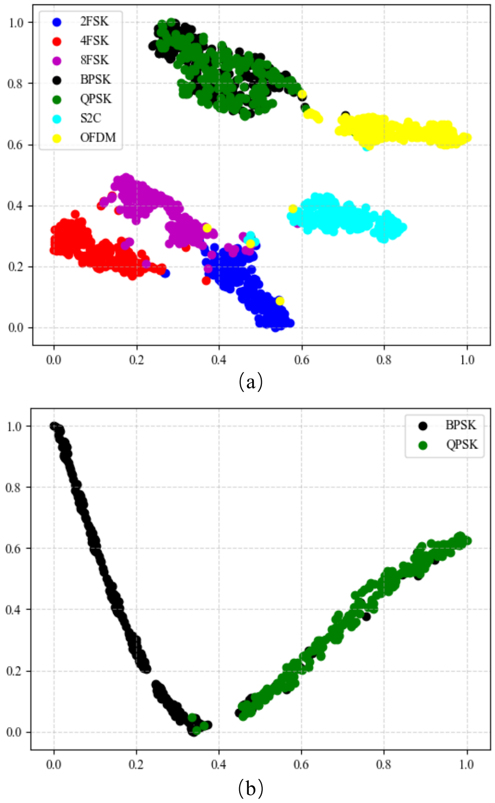


FIGURE 19. T-SNE plots for the signal features extracted by (a) Att-CNN module; (b) SAE module.

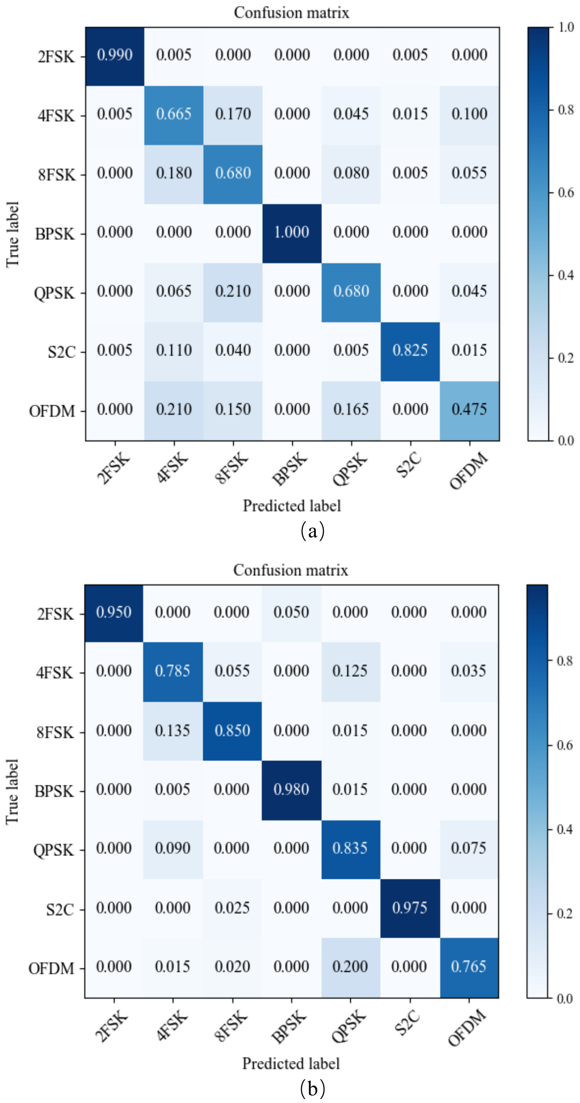


FIGURE 20. Confusion matrices of different methods in the Wuyuan Bay experiment. Note that the vertical and horizontal coordinates represent the true and predicted labels, respectively. (a) SAE-2048 method: The classification accuracies reach over 82.5% for 2FSK, BPSK and S2C, while 4FSK, 8FSK, QPSK, and OFDM signals have poor classification accuracies of 66.5%, 68%, 68% and 47.5%, respectively; (b) Proposed method: The classification accuracies reach over 83.5% for most signals, except when 4FSK and OFDM have relatively low accuracies of 78.5% and 76.5%.

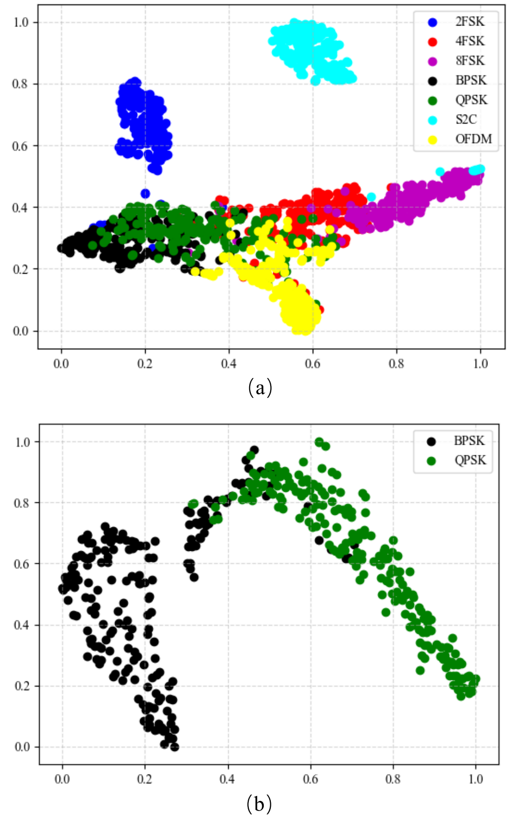
of [1.16, 1.95]. The STFT diagrams of the 4FSK signals collected from Wuyuan Bay show evident inter-symbol interferences between the different symbols. It is difficult to clearly distinguish the boundary of the symbols. This indicates that the received signals are seriously influenced by the multi-path effect of the actual UWA channel.

Finally, after impulsive noise preprocessing, burst detection and dimension preprocessing, a small training set with 50 signal examples per modulation and a testing set with 200 examples per modulation are obtained in each water region. Each signal example has a short duration of 170.7 ms (i.e., 8192 samples). Based on the limited training data, the proposed method and the SAE-2048 method were adopted to recognize the testing signal examples. The transfer training set in our method is built under channel h_0. Fig. 18 compares the classification confusion matrices of the two methods in the artificial lake experiment.

As shown in Fig. 18, the SAE-2048 method has a poor performance for several types of modulations when the training data are limited and the testing signal duration is short. In comparison, our method performs significantly better under these unfavorable conditions. The classification accuracies reach over 80% for most signals, except when few 8FSK examples are misjudged as S2C-modulated. In fact, the adoption of the transfer learning strategy in our method has significantly alleviated the demand for training data from the testing channel. Our method does not require power spectra estimation, thus reducing the need for the number of transmitted symbols within a single testing example. The results preliminarily prove the effectiveness of the proposed method under actual UWA channels.

To vividly demonstrate this, a t-SNE technique [44] is adopted to visualize the extracted signal features of the proposed Att-CNN and SAE module. The output of their penultimate layers is mapped to a 2D plane through dimension reduction. Fig. 19 shows the visualization results of the Att-CNN and SAE module.

As shown in Fig. 19(a), the Att-CNN module can effectively extract the features for most signals and divide them into several clusters, except when those of BPSK and QPSK



(a)

(b)

FIGURE 21. **T-SNE plots for the signal features extracted by (a) Att-CNN module; (b) SAE module.**

signals are completely confused. Fortunately, Fig. 19(b) shows that the two clusters of them can be effectively distinguished by the SAE module, though few features are still confused. Overall, the visualization results shown in Fig. 19 are consistent with the classification results presented in Fig. 18(b).

When it comes to the Wuyuan Bay experiment, the confusion matrices of the proposed method and the SAE-2048 method are obtained and shown in Fig. 20. Fig. 21 shows the feature visualization results.

As shown in Fig. 20, our method still outperforms the SAE-2048 method in this experiment. The classification accuracies are over 80% for most signals, except when few 4FSK and OFDM examples are misjudged as QPSK-modulated. Moreover, the visualization results in Fig. 21 are similar to those in Fig. 19 and are consistent with the classification results shown in Fig. 20(b). Hence, the above results have further proved the effectiveness of the proposed method in actual marine environments.

IV. CONCLUSION

In this study, we developed a novel HNN-based AMC method for UWA communication signals. The proposed Att-CNN and SAE module are combined to effectively extract the temporal and spectral features of signals. Moreover, problems, such as short signal duration and data scarcity, are resolved by adopting a DR approach and a transfer learning strategy, respectively. The results of simulation experiments and practical signal tests both demonstrated that the proposed method is robust against UWA channels and ambient noise, with improved performance.

The proposed method suggests the following research directions. First, the network structure can be further improved to reduce the network parameters, as well as reducing the requirement for training data. Second, more signal modalities, such as the cycle spectra can be used to enhance the robustness against unknown marine environments. Third, other fusion mechanism, such as the early fusion can be tried to improve the overall performance based on more modalities. Finally, the case when no training data from the testing channel is available should also be considered.

REFERENCES

[1] C. Zhang, Y. Wang, and X. Guan, "Chaotic modulation detection for underwater acoustic communications via instantaneous features," in *Proc. OCEANS MTS/IEEE Monterey*, Sep. 2016, pp. 1–5.

[2] Q. Zhou, H. Sun, and M. Zhou, "One method of standard recognition of underwater acoustic signal," *Commun. Countermeas.*, vol. 36, no. 2, pp. 12–17, 2017.

[3] Z. Wu, T. Yang, Z. Liu, and V. Chakarvarthy, "Modulation detection of underwater acoustic communication signals through cyclostationary analysis," in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, Oct. 2012, pp. 1–6.

[4] J. Sanderson, X. Li, Z. Liu, and Z. Wu, "Hierarchical blind modulation classification for underwater acoustic communication signal via cyclostationary and maximal likelihood analysis," in *Proc. IEEE Mil. Commun. Conf. MILCOM*, Nov. 2013, pp. 29–34.

[5] X. Li, Q. Han, Z. Liu, and Z. Wu, "Novel modulation detection scheme for underwater acoustic communication signal through short-time detailed cyclostationary features," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2014, pp. 624–629.

[6] Y. Ge, Z. Ye, Q. Zhou, C. Lin, and J. Qi, "Modulation recognition for underwater communication signals in alpha stable distribution noise," *Commun. Countermeas.*, vol. 35, no. 2, pp. 16–20, Jun. 2016.

[7] X. Zhang, S. Anwar, N. U. R. Junejo, H. Sun, Q. Jie, and C. Lin, "Application of cyclic cumulant in recognition of underwater communication system," in *Proc. IEEE Int. Conf. Signal Process., Commun. Comput. (ICSPCC)*, Oct. 2017, pp. 1–4.

[8] N. Alyaoui, A. Kachouri, and M. Samet, "Automatic modulation classification for underwater acoustic Communications," in *Proc. 10th IEEE Int. Symp. Signal Process. Inf. Technol.*, Dec. 2010, pp. 166–170.

[9] E. Cheng, J. Yan, H. Sun, and J. Qi, "Research on MPSK modulation classification of underwater acoustic communication signals," in *Proc. IEEE/OES China Ocean Acoust. (COA)*, Jan. 2016, pp. 1–5.

[10] Y. Ge, X. Zhang, and Q. Zhou, "Modulation recognition of underwater acoustic communication signals based on joint feature extraction," in *Proc. IEEE Int. Conf. Signal, Inf. Data Process. (ICSIDP)*, Dec. 2019, pp. 1–4.

[11] W. Jiang, X. Cao, and F. Tong, "Modulation recognition method of underwater acoustic communication signals using SVM," *J. Xiamen Univ. Sci.*, vol. 54, no. 4, pp. 534–539, Jan. 2015.

[12] W. Jiang, F. Tong, Y. Dong, and G. Zhang, "Modulation recognition of non-cooperation underwater acoustic communication signals using principal component analysis," *Appl. Acoust.*, vol. 138, pp. 209–215, Sep. 2018.

[13] H. Li, Y. Cheng, W. Dai, and Z. Li, "A method based on wavelet packets-fractal and SVM for underwater acoustic signals recognition," in *Proc. 12th Int. Conf. Signal Process. (ICSP)*, Jan. 2015, pp. 2169–2173.

[14] Y. Lu, X. Wang, P. Zhao, and H. Zhou, "Identifications of underwater acoustic communication signals classification based on time-frequency analysis and neural Network," *Ence Technol. Rev.*, vol. 29, no. 28, pp. 33–36, Sep. 2011.

[15] Z. Zhao, W. Shilian, W. Zhang, and Y. Xie, "A novel automatic modulation classification method based on stockwell-transform and energy entropy for underwater acoustic signals," in *Proc. IEEE Int. Conf. Signal Process., Commun. Comput. (ICSPCC)*, Aug. 2016, pp. 1–6.

[16] C. N. Marcoux, B. Chandna, and B. J. Blair, ''Blind equalization and automatic modulation classification of underwater acoustic signals,'' *J. Acoust. Soc. Amer.*, vol. 144, no. 3, p. 1729, Sep. 2018.

[17] D. Li-Da, W. Shi-Lian, and Z. Wei, ''Modulation classification of underwater acoustic communication signals based on deep learning,'' in *Proc. OCEANS MTS/IEEE Kobe Techno-Oceans (OTO)*, May 2018, pp. 1–4.

[18] Q. Zhou, M. Shao, C. Li, J. Yin, and X. Han, ''Underwater communication system recognition with deep learning,'' *Tech. Acoust.*, vol. 36, no. 5, pp. 425–426, Oct. 2017.

[19] H. Yang, S. Shen, J. Xiong, and X. Zhang, ''Modulation recognition of underwater acoustic communication signals based on denoting & deep sparse autoencoder,'' in *Proc. INTER-NOISE NOISE-CON Congr. Conf.*, Aug. 2016, pp. 5144–5149.

[20] X. Yu, L. Li, J. Yin, M. Shao, and X. Han, ''Modulation pattern recognition of non-cooperative underwater acoustic communication signals based on LSTM network,'' in *Proc. IEEE Int. Conf. Signal, Inf. Data Process. (ICSIDP)*, Dec. 2019, pp. 1–5.

[21] N. Jiang and B. Wang, ''Underwater communication signals' modulation recognition based on sparse autoencoding network,'' *J. Signal Process.*, vol. 35, no. 1, pp. 103–114, Jan. 2019.

[22] Y. Li, B. Wang, and G. Shao, ''A method of modulation recognition of underwater acoustic communication signals based on Alexnet,'' *Tech. Acoust.*, vol. 37, no. 6, pp. 1–2, Dec. 2018.

[23] X. Yao, H. Yang, and Y. Li, ''Modulation recognition of underwater acoustic communication signals based on convolutional neural networks,'' *Unmanned Syst. Technol.*, vol. 1, no. 4, pp. 68–74, Apr. 2018.

[24] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, ''Multimodal fusion for multimedia analysis: A survey,'' *Multimedia Syst.*, vol. 16, no. 6, pp. 345–379, Apr. 2010.

[25] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, ''Multimodal machine learning: A survey and taxonomy,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[26] G. Aceto, D. Ciuonzo, A. Montieri, and A. Pescapè, ''MIMETIC: Mobile encrypted traffic classification using multimodal deep learning,'' *Comput. Netw.*, vol. 165, pp. 1–14, Oct. 2019.

[27] G. Aceto, D. Ciuonzo, A. Montieri, and A. Pescapé, ''Toward effective mobile encrypted traffic classification through deep learning,'' *Neurocomputing*, vol. 409, pp. 306–315, Oct. 2020.

[28] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, ''ModDrop: Adaptive multi-modal gesture recognition,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1692–1706, Aug. 2016.

[29] K. G. Kebkal and R. Bannasch, ''Sweep-spread carrier for underwater communication over acoustic channels with strong multipath propagation,'' *J. Acoust. Soc. Amer.*, vol. 112, no. 5, pp. 2043–2052, Nov. 2002.

[30] J. G. Veitch and A. R. Wilks, ''A characterization of arctic undersea noise,'' *J. Acoust. Soc. Amer.*, vol. 77, no. 3, pp. 989–999, Mar. 1985.

[31] M. Chitre, J. Potter, and O. S. Heng, ''Underwater acoustic channel characterisation for medium-range shallow water communications,'' in *Proc. Oceans MTS/IEEE Techno-Ocean*, Mar. 2004, pp. 40–45.

[32] B. Hu and D. Yang, ''Symmetic alpha-stable distributions for analysis of underwater impulsive noise,'' *Tech. Acoust.*, vol. 25, no. 2, pp. 134–139, Apr. 2006.

[33] G. Samorodnitsky and M. Taqqu, ''Stable non-gaussian random processes: Stochastic models with infinite variance,'' *J. Am. Stat. Assoc.*, vol. 90, pp. 805–806, Jau. 1996.

[34] R. Barazideh, W. Sun, B. Natarajan, A. V. Nikitin, and Z. Wang, ''Impulsive noise mitigation in underwater acoustic communication systems: Experimental studies,'' in *Proc. IEEE 9th Annu. Comput. Commun. Workshop Conf. (CCWC)*, Mar. 2019, pp. 880–885.

[35] Y. Li, B. Wang, G. Shao, S. Shao, and X. Pei, ''Blind detection of underwater acoustic communication signals based on deep learning,'' *IEEE Access*, vol. 8, pp. 204114–204131, Nov. 2020.

[36] S. Zheng, P. Qi, S. Chen, and X. Yang, ''Fusion methods for CNN-based automatic nodulation classification,'' *IEEE Access*, vol. 7, pp. 66496–66504, 2019.

[37] J. Hu, L. Shen, and G. Sun, ''Squeeze-and-excitation networks,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7132–7141.

[38] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, and U. Montreal, ''Greedy layer-wise training of deep networks,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 19, Jan. 2007, pp. 153–160.

[39] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, ''How transferable are features in deep neural networks?'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, Nov. 2014, pp. 3320–3328.

[40] D. P. Kingma and J. L. Ba, ''Adam: A method for stochastic optimization,'' in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2015, pp. 1–15.

[41] Z. Chuangzhan and J. Xin, ''Modulation recognition method of communication signals based on correlation characteristics,'' in *Proc. IEEE Int. Conf. Signal Process., Commun. Comput. (ICSPCC)*, Sep. 2015, pp. 1–5.

[42] A. Zhang, T. Qiu, and X. Zhang, ''A new underwater acoustic signals processing approach to α-stable distribution,'' *J. Electron. Inf. Technol.*, vol. 27, no. 8, pp. 1201–1204, Aug. 2005.

[43] J. H. McCulloch, ''Simple consistent estimators of stable distribution parameters,'' *Commun. Stat. Simul. Comput.*, vol. 15, no. 4, pp. 1109–1136, Jan. 1986.

[44] A. Karpathy, J. Johnson, and L. Fei-Fei, ''Visualizing and understanding recurrent networks,'' Jun. 2015, *arXiv:1506.02078*. [Online]. Available: http://arxiv.org/abs/1506.02078

YONGBIN LI was born in 1996. He received the B.S. degree from PLA Information Engineering University, in 2018, where he is currently pursuing the master's degree. His research interest includes underwater acoustic communication signal processing.

BIN WANG received the Ph.D. degree from PLA Information Engineering University, in 2007. Her research interest includes underwater acoustic communication signal processing.

GAOPING SHAO received the Ph.D. degree from the Beijing Institute of Technology, in 2009. His research interest includes communication signal processing.

SHUAI SHAO received the M.S. degree from the University of Central Lancashire, in 2014. His research interests include digital signal and image processing.