frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 12 March 2024
DOI 10.3389/fmars.2024.1331717

OPEN ACCESS

EDITED BY
Huiyu Zhou,
University of Leicester, United Kingdom

REVIEWED BY
Weifeng Sun,
China University of Petroleum (East China), China
Rui Qin,
Manchester Metropolitan University, United Kingdom
Corentin Houpert,
University of Leicester, United Kingdom

*CORRESPONDENCE
Xiao Cheng
chengxiao_ouc@163.com

RECEIVED 01 November 2023
ACCEPTED 21 February 2024
PUBLISHED 12 March 2024

CITATION
Wang Y, Xiao J, Cheng X, Wei Q and Tang N (2024) Underwater acoustic signal classification based on a spatial–temporal fusion neural network.
Front. Mar. Sci. 11:1331717.
doi: 10.3389/fmars.2024.1331717

Underwater acoustic signal classification based on a spatial–temporal fusion neural network

Yan Wang[1], Jing Xiao[1], Xiao Cheng[1]*, Qiang Wei[1] and Ning Tang[2]

[1]School of Physics and Electronic Engineering, Taishan University, Tai'an, China, [2]College of Electronic Engineering, Faculty of Information Science and Engineering, Ocean University of China, Qingdao, China

In this paper, a novel fusion network for automatic modulation classification (AMC) is proposed in underwater acoustic communication, which consists of a Transformer and depth-wise convolution (DWC) network. Transformer breaks the limitation of sequential signal input and establishes the connection between different modulations in a parallel manner. Its attention mechanism can improve the modulation recognition ability by focusing on the key information. DWC is regularly inserted in the Transformer network to constitute a spatial–temporal structure, which can enhance the classification results at lower signal-to-noise ratios (SNRs). The proposed method can obtain more deep features of underwater acoustic signals. The experiment results achieve an average of 92.1% at −4 dB ≤ SNR ≤ 0 dB, which exceed other state-of-the-art neural networks.

KEYWORDS

underwater acoustic communication, modulation classification, signal recognition, deep learning, neural network

1 Introduction

In wireless communication, AMC is one of the important tools for signal detection, and it plays an irreplaceable role in the civil and military field (Chithaluru et al., 2022; Liu et al., 2022; Teekaraman et al., 2023; Zhang et al., 2023). In the civil field, AMC is the basis for channel parameter estimation and spectrum monitoring. In the military field, AMC is widely used in information interception, interference selection, and radiation source classification, among others. However, the wireless environment of underwater communication is complex and changeable, and the modulation types become more and more diverse, all of which bring great challenges to the acoustic signal of modulation identification technology. Therefore, it is urgent to study efficient and intelligent AMC methods in underwater acoustic communication transport.

In underwater acoustic signal modulation recognition, the challenge arises from discerning intricate signal characteristics, which varies environmental noises and

disturbances. Unlike terrestrial communication, underwater channels are subject to unique interference such as multipath propagation, varying sound speed profiles, and ambient noises from natural and anthropogenic sources. These factors not only obscure signal clarity but also introduce variability and unpredictability in signal behavior, complicating the modulation recognition process.

The dilemma intensifies with the diverse modulation schemes used in underwater communication, each characterized by distinct features that necessitate precise identification and classification Addressing these complexities necessitates the development of sophisticated AMC methods, which are adept at handling the dynamic nature of underwater signals. It is efficient in differentiating between an array of modulation types under varying channel conditions.

In a real underwater communication environment, there are many factors that affect acoustic signal transmission, such as ocean physics movement, maritime commercial activities, and marine organisms (Cai et al., 2022; Zhang Y. et al., 2022; Zhai et al., 2023; Zheng T. et al., 2023). The underwater acoustic channel is often more complex and changeable than the land air channel. There are two categories in AMC technology: likelihood-based (LB) methods and feature-based (FB) methods (Hamee and Wadi, 2015; Abu-Romoh et al., 2018; Li et al., 2019; Hreshee, 2020). In the received signal modulations, the likelihood function of unknown parameters can be maximized by LB methods, which change the identification problem into multiple hypothesis testing. There is the most likely outcome by the likelihood ratio, and it requires a lot of prior knowledge of transmission environment. The suboptimal performance is obtained by FB methods, which have huge advantages in computational complexity and robustness. Deep learning algorithms (DLAs) are one of the important branches of FB, which can automatically extract features and produce classifying results.

DLAs have made remarkable strides in image classification, natural language processing, and speech recognition (Dong S. et al., 2021; Song et al., 2022; Menghani, 2023; Xu et al., 2023; Zheng C. et al., 2023). There are a wealth of studies in underwater AMC. In Wang et al. (2019), the convolutional neural network (CNN) is the efficient extractor in spatial domain, and the discriminative information of signal features is fully obtained. The underwater channel interference mixed with oceanic noise can be effectively mitigated to improve the recognition performance. The imbalanced class of underwater acoustic modulations is studied in Dong Y. et al. (2021), and the redesigned loss function can significantly stress the recognition effects, which assume the exponential categorical cross entropy in CNN. In the few shot scene, Wang et al. (2022a) shows that the underwater dataset containing the impulse noise can achieve better results by the employed network, which adopts the attention mechanism in the network design. Wang et al. (2022b) uses the hybrid network to identify the modulation styles in multiple underwater signal receiving devices. The method can not only obtain multi-channel signal features, but also better classify these features, which can greatly emphasize the recognition effects. There is the redesigned autoencoder of extractor in Huang et al. (2022), which employs the K-nearest neighbors method to classify features. While the recognition rate is impressively improved, the recognition time is also shorter. It is due to the fact that there are acquired high-quality features and the appropriate recognizing method. In Gao et al. (2022), the underwater acoustic modulation data are collected in three different scenarios, which are applied to the comparative learning. More significant discriminative information of modulation styles are earned on supervised conditions, which can efficiently differentiate between MPSK and MFSK at low signal-to-noise ratios (SNRs). Zhang W. et al. (2022) adopt two neural network forms, which consist of a CNN and a recurrent neural network (RNN). The classification results are improved by the wider network structure. There is a shorter recognition time to employ the 1D convolution kernel and remove the pooling operation. In Fang et al. (2022), the wavelet transformation augments the acoustic signal data to weaken the underwater communication interference. The redesigned random forest (RF) enhances the recognition effect with the assistance of signal spectrum characteristics, which optimizes computational complexity.

DLA solutions have made a lot of recent progress in the modulation classification field. There are still many shortcomings in the existing methods. First of all, the structure of CNN methods is often complex to extract more deep features, which is prone to the overfitting problem resulting in poor practical outcome. The maximum number of Vapnik–Chervonenkis dimensions that the large-scale CNN model can classify training samples is too high. It leads to fitting noise and unrepresentative features in training samples, which makes the model unable to really categorize the true distribution of the whole data. Secondly, it is difficult for the neural network based on RNN, in the main form of long short-term memory (LSTM), to solve the gradient problem after superposition, which is almost impossible to further grow in the recognition effect. In the underwater modulation classification, it is important to effectively obtain the hidden signal information, and the two network forms cannot cope with the underwater acoustic interference, such as obvious multi-path effect, serious time delay of acoustic signal propagation, and marine ambient noise, which leads to serious signal attenuation and modulation constellation confusion. Aiming at the problems of CNN and RNN, an autoencoder is adopted, which is different from the two neural network architectures. Its implementation approach is the Transformer architecture (Dosovitskiy et al., 2020), which demonstrates fairly high accuracy in underwater acoustic modulation recognition. Transformer is essentially a network architecture of an encoder–decoder structure, which consists of the self-attention mechanism and the feed forward neural network of multilayer perceptron (MLP).

In this paper, the innovative approach addresses these challenges by combining advanced spatial–temporal fusion techniques with a neural network architecture, which is tailored for underwater acoustic signal modulation recognition. The proposed method leverages a unique combination of spatial domain analysis and temporal sequence modeling, enabling it to effectively handle the complex characteristics of underwater signals.

By integrating these techniques, the proposed solution not only mitigates the issues of signal attenuation and modulation constellation confusion but also overcomes the limitations of traditional CNN and RNN models, including overfitting and gradient problems. Furthermore, the proposed approach combines spatial–temporal fusion techniques with a neural network architecture tailored for this purpose. It effectively handles complex signal characteristics and mitigates issues like signal attenuation and modulation constellation confusion, which overcomes the limitations of traditional models. Additionally, it is designed to filter out noise and irrelevant features, capturing the true data distribution. These capabilities lead to more accurate and robust modulation recognition, highlighting the method's potential for significant improvements in both accuracy and practical applicability for underwater acoustic signal classification. Compared to using DWC alone (Hu et al., 2021), our method enhances classification in lower SNRs by integrating Transformer with DWC, utilizing advanced spatial–temporal fusion techniques for improved accuracy and applicability in underwater acoustic signal classification.

The contributions are summarized as follows:

1. The Transformer network is introduced to handle long temporal signal series, and the high-dimensional features of temporal domain signals are dynamically acquired in the long-range sequence correlation, which enhances the recognition ability at lower SNRs.
2. A novel attention mechanism is proposed that is efficient and effective at finding a small number of important differentiate information from weak underwater acoustic signals. This mechanism is able to model pairwise attention over a longer temporal signal.
3. Multiple DWC blocks are creatively inserted to the stacked Transformer modules. By employing the fusion network, the model gains the ability of extracting spatial representations of underwater signal characteristics. Meanwhile, it has a lower computing burden and better classification performance.

2 The proposed method

2.1 Signal model

The AMC task is effectively a multi-classification problem, which bears a striking resemblance to other conventional work in the DLA field. The form of received signals is the complex representation in the temporal domain, which includes all modulation styles. The underwater received acoustic signals (Equation 1) can be represented as:

$$s(t) = c(t) * h(t;\tau) + n(t) \tag{1}$$

where $s(t)$ is the received transmission signal, $c(t)$ is the emitted complex modulation signal, $*$ is the signal convolution operation, $n(t)$ is the additive Gaussian white noise, $h(t;\tau)$ is the underwater acoustic channel impulse response, τ is the signal time delay, and t is the signal time.

2.2 Proposed network

2.2.1 Transformer model

The proposed network architecture is shown in Figure 1. Let V be an input signal tensor of dimensions $H \times W \times C$, where $H \times W$ correspond to the temporal dimensions of the input, and C represents the number of channels. For the input V, the goal is to learn a set of K abstract features denoted as $U = [u_i]_{i=1}^{K}$. To elaborate further, $u_i = B_i(V)$ serves as a constructed feature function that maps the input V to a feature vector $u_i : \mathbb{R}^{H\times W\times C} \mapsto \mathbb{R}^{C}$. During this process, K feature functions B_i are learned, enabling adaptive selection of latent information from V. Consequently, the abstract features obtained in this manner do not constitute a fixed partition of the input tensor, which represent a set of adaptively changing identification selections. Each input will mine different high-level representations with high-level features. It is possible to model the long-range dependencies and interaction distinction information of signals in special underwater noise environments.

To control the computational cost of the model, K is set to be less than both H and W. This can significantly reduce the

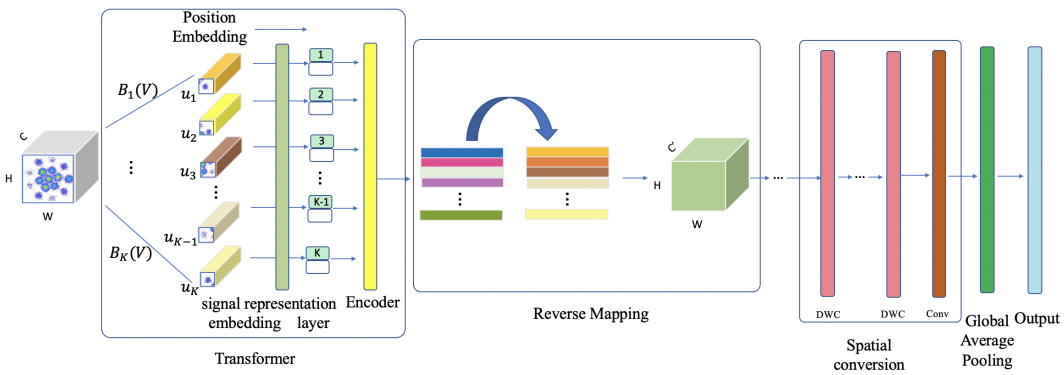

FIGURE 1
The architecture of the proposed network.

computational complexity required by subsequent modules in the model, ensuring the training speed and parameter capacity of the model. In this context, the ith feature $u_i = B_i(V)$ are implemented using a temporal attention mechanism. In the learning process of the model, it is necessary to generate a computational weight map corresponding to the size of $H \times W$ according to the input V. These weight maps are element-wise multiplied with V to create abstract data combinations with temporal features. Specifically, let $\mu_i(V)$ represent the function responsible for generating $H \times W \times 1$ weight maps. Each feature u_i (Equation 2) is generated as follows:

$$u_i = B_i(V) = GAP(V \odot B_{i\_}) = GAP(V \odot \eta(\mu_i(V))) \tag{2}$$

where $\odot$ signifies element-wise multiplication, and $B_{i\_} \in \mathbb{R}^{H\times W\times C}$ is an intermediate weight tensor calculated by the function $\mu_i(V)$ and the broadcasting function $\eta(\cdot)$. Finally, temporal global average pooling $GAP(\cdot)$ is applied to reduce the dimensions to $\mathbb{R}^C$. The resultant feature aggregations are aggregated to form the output tensor $U = [u_i]_{i=1}^K \in \mathbb{R}^{K\times C}$. The entire process takes the form of element-wise temporal self-attention. The functions $u_i(\cdot)_{i=1}^K$ are collectively implemented as a single or a series of convolutional layers with a channel size of K, followed by the sigmoid function, facilitating the generation of results.

The underwater acoustic signals are preprocessed, and each input is treated as an element, which equates a block operation. The dimension is reduced by the liner projection of flattened spots, and a liner embedding has been obtained. In this way, the original underwater acoustic signals are serialized, and there are a set of preprocessing features and position embedding. The position embedding contains the position information in the sequence, which is regarded as trainable input variables. The preprocessed signal data are transmitted to the Encoder as the input for the deep feature extraction. The proposed Encoder structure is shown in Figure 2. The process (Equations 3–5) can be expressed as:

$$y'_\ell = MHA(LN(y_{\ell-1})) + y_{\ell-1} \tag{3}$$

$$y_\ell = MLP(LN(y'_\ell)) + y'_\ell \tag{4}$$

$$y = LN(y_\ell) \tag{5}$$

$MHA(\cdot)$ is the multi-head attention mechanism, and $LN(\cdot)$ is layer normalization. $y_{\ell-1}$ represents the $(\ell-1)$-th layer input, and y_ℓ' is the multi-head attention output in the ℓ-th layer; $MLP(\cdot)$ is the multilayer perceptron, and y is the final output of the Encoder.

y is then sent to the reverse mapping block where a reverse attention mechanism is employed to provide an adaptive fusion, enabling selective reconstruction of the output from the labels based on content and context. A reverse mapping function F (Equation 6) is learned to map the label tensor back to the original shape:

$$V' = F(y) \in \mathbb{R}^{H\times W\times C'} \tag{6}$$

where $\mathbb{R}^{C'}$ is typically different from the original input channel C. In other words, the proposed network initially learns a reverse attention mechanism that generates attention vectors $\alpha_{\iota\varepsilon}(y)$ of shape $\mathbb{R}^{K\times 1}$ for each position (ι,ε). These attention vectors (Equation 7) are multiplied element-wise with the reshaped labels y, and a weighted sum is performed to generate the output for each position (ι,ε),

$$V'_{\iota\varepsilon} = \sum_{i=1}^{K} \alpha_{\iota,\varepsilon}(y)_i \tag{7}$$

where $y_i \in \mathbb{R}^{C'}$ represents the vector representation of the ith label. The above equation can be written in matrix form (Equation 8):

$$V' = \mathcal{F}(\alpha(y), y) \tag{8}$$

where $\alpha(y) \in \mathbb{R}^{H\times W\times K}$ denotes the attention vectors for all positions (ι,ε), and $\mathcal{F}$ represents matrix multiplication and summation operations. The distinction in dimensions, $\mathbb{R}^{K\times 1}$ for attention vectors and $\mathbb{R}^{C'}$ for label vectors, underscores the adaptive fusion capability of our reverse attention mechanism, enabling precise context-aware reconstruction from label information to the desired output shape, enhancing the model's interpretability and effectiveness in handling complex signal characteristics.

The signal representation embedding layer is integrated by the intermediate abstract multimodal fusion layer, and the complete temporal Transformer is constructed. The input V first goes through the signal representation embedding layer module to generate markers U, which are then fed into a standard Transformer encoder, and finally restored back to the original shape via the intermediate abstract multi-modal fusion layer module. By stacking multiple layers of the temporal Transformer, a more powerful model can be built. The temporal Transformer retains the advantages of Transformer in modeling, while adapting computations to the dimensions through labeling and fusion operations. This provides an efficient and flexible method for modeling interactions in underwater signal recognition, improving the performance and deployment practicality of network models.

2.2.2 Transformer embedding CNN

The relative position information refers to the fact that the signal distribution position is used to distinguish the modulation categories in the underwater acoustic signal modulation constellation. The original underwater acoustic data do not contain the relative position information of modulations, leading to the same effect in a different position vector. The different position vector corresponds to the position information contained in the input underwater acoustic signal sequence, which is input to the Transformer network as a vector. It is difficult to distinguish the modulation types in the spatial dimension. The discrimination ability of Transformer can be effectively improved by the position information. The attention mechanism of Transformer can remember the key distinguishing information like the human visual attention mechanism.

In general, CNN gives the same weight to all the position information, which will limit the expression ability of the model. It is almost impossible to distinguish the modulation types that are seriously disturbed by the underwater environment. The Transformer attention mechanism is used for the feature aggregation, which can adaptively adjust the weight of feature

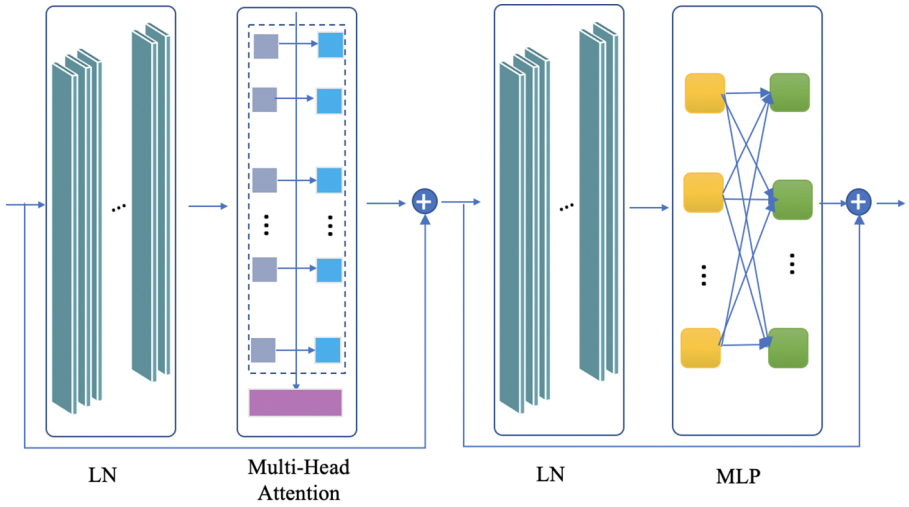

FIGURE 2
The structure of the proposed Encoder architecture.

aggregation according to the relationship between the underwater acoustic signal sequence and the location information. The model is improved to alleviate the signal fading, which enhances the modulation recognition ability.

If there are no strong constraints in the training of Transformer, the recognition ability of CNN is better than that of the same size model on smaller dataset. Compared to CNN, Transformer has less prior knowledge of inductive bias. Unlike CNN, Transformer can learn by itself from data. In the absence of enough data to pre-train, it is impossible to get a good transfer learning effect on downstream tasks, and the Transformer obtains similar or better results than the current best CNN. Therefore, the better way is combining Transformer and CNN in network design. Transformer has the relatively strong global modeling ability, which can acquire the key classification information at a lower SNR. CNN has an inductive bias ability that can effectively improve the feature extraction ability on a smaller-scale dataset.

CNN is regularly inserted in the Transformer architecture, and there is the spatial conversion of the underwater acoustic signal. The combination of modulation classification information can be adaptively selected. The tensors extracted by the input middle layer can simulate the modulation spatial–temporal relationship of underwater acoustic signals in the adaptive splitting form, and the recognition results can be promoted at the lower SNR. The process (Equations 9–11) can be expressed as:

$$z_j = DWC_{Gelu}(V') \tag{9}$$

$$Z = Conv_{Sigmoid}(\Gamma_{j=1}^{J} z_j) \tag{10}$$

$$O = \lambda(V' \odot Z) \tag{11}$$

The spatial conversion layers of the inserted CNN is the DWC form in Figure 3. $DWC_{Gelu}(\cdot)$ is the depth-wise convolution with the Gelu activation function. The expression $H \times W \times C'$ is split into $H \times$

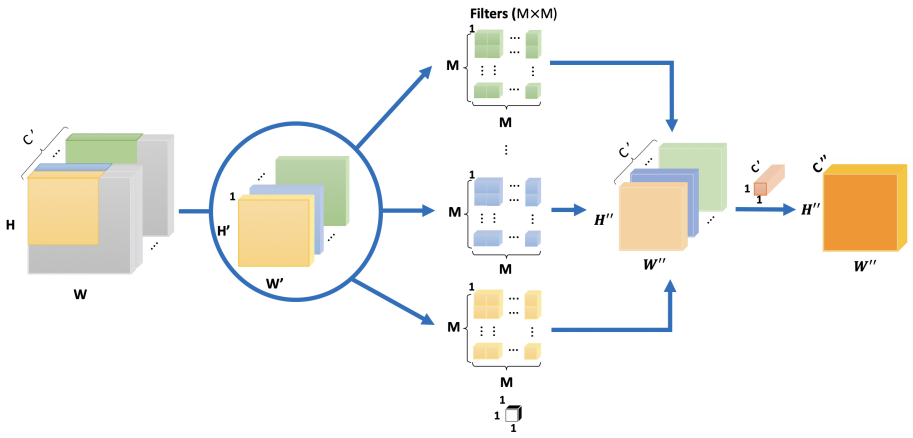

FIGURE 3
DWC convolution operation process.

$W \times 1$, which means that a tensor with dimensions of height H, width W, and depth C' is divided into smaller tensors, each having dimensions of height H, width W, and depth 1. Following this, a convolution operation with a kernel size of $M \times M \times 1$ is applied, resulting in a tensor with dimensions $H'' \times W'' \times C'$. Subsequently, another convolution operation with a kernel size of $1 \times 1 \times C'$ is applied, which yields the final output tensor of depth C'' and dimensions $H'' \times W'' \times C''$. $Conv_{Sigmoid}(\cdot)$ is the two-dimension convolution with the Sigmoid activation function, $\Gamma_{j=1}^{J}[\cdot]$ is the number of layers of DWC, j is the corresponding layer, $j = 1,\ldots,J$. Through an optimized spatial transformation that preserves essential feature information, Z is the weight map of the spatially transformed output, $\odot$ is the element-wise multiplication, $\lambda(\cdot)$ is the global pooling. The method not only optimizes the computational efficiency, but also separately convolves the exacted features in space, which is more conducive to distinguishing the detailed features of the underwater acoustic signals to strengthen the recognition ability.

3 Experiment

In the underwater acoustic wireless channel (Wang et al., 2022a), the generated signals are more approximate to the realistic situation of disturbances. The dataset involves 10 types of modulation signals, which is BPSK, QPSK, 8PSK, 4PAM, 16QAM, 64QAM, FM, DSB, CPFSK, and 4FSK. Other setting experiment parameters are shown in the Table 1.

A total of 2,000,000 modulation data are included in the dataset, and SNRs are in the range of −20 dB to +18 dB. The dataset is divided into a training set, a validation set, and a testing set with 60%, 20%, and 20%, respectively. There is a complex floating point I/Q value in each signal data, which has a length of 256. Otherwise, the time fading model is Rayleigh distribution. The range of SNRs is from −20 dB to 20 dB. Gaussian white noise is considered for the additive noise, which is bandlimited and has a zero mean. The random number generator seed of the noise source is set to 0x1498.

When utilizing a patch size of 16 × 16, a batch size of 64, and constructing the model with two Transformer encoders and three DWC blocks, the classification accuracy of the model achieves an optimal performance. This specific combination of hyperparameters demonstrates the model's ability to efficiently capture complex acoustic patterns and spatial features present in the data. These results underscore the significance of our fusion model settings for achieving superior classification outcomes.

TABLE 1 The experiment parameters.

Parameters	Data	Comment
Doppler shift carrier Frequency	5×10^3 10 kHz	
Symbol transmission rate	1,000 symbol/s	
Sample rate of the maximum deviation	15 Hz	In the random mode
Standard offset drift process	1 Hz per sample	In the sample rate
Time fading model	Rayleigh distribution	
Frequency selective fading	20 cosines	
Filter of the raised cosine pulse shaping	0.35 roll-off factor	

The classification results of the proposed method are shown in Figure 4. When SNR is less than −10 dB, the classification rate is lower than 49.8%. With the increase of SNRs, the recognition accuracy is obviously improved in the magnitude. The classification rate reaches nearly 93.7% at SNR = −2 dB. The proposed method can realize the effective identification of various modulation styles at the lower SNR. At SNR ≥ 0 dB, the recognition effect continues to grow, and the classification results can achieve approximately 97.9% on average. The results demonstrate the effectiveness of the proposed method in the network structure design.

In Figure 5, the classification rate of DWC is higher than CNN, SeparableCNN, and TransposeCNN. Compared with three CNN forms, DWC has obvious advantages in recognition accuracy at lower SNRs. In particular, the recognition effect of DWC is nearly 16.9% stronger, on average, than that of CNN at the range from −8 dB to −4 dB in Figure 5A. At the same SNR range, the similar situation also appears in Figures 5B, C, and DWC is approximately 7.8%, 5.8% higher than SeparableCNN and TransposeCNN on average, respectively. The convolution form of DWC resists the influence of underwater transmission environment, which can achieve higher classification results.

At a certain SNR, the classification result in varieties of modulation types is shown in Figure 6 at bit group = 256. At SNR = −6 dB, 4FSK and 8FSK are misidentified as SSB in Figure 6A. At a lower SNR, the analog signal waveforms are easily confused to lead to the bad results. Simultaneously, 16QAM and 32QAM are poorly identified, and there is a similar constellation diagram to the two modulation types to cause the bad results. When SNR is improved to −4 dB, in Figure 6B, the classification results of 4FSK and 8FSK are greatly enhanced. The classification rate for 4FSK improved significantly by 62%, with misidentifications as SSB and 8FSK reduced to 13% and 12%, respectively. Additionally, the recognition rates for 8PSK and SSB also improved by 10% and 4%. At the SNR, the network's capability to discern between different modulation signatures was enhanced, leading to significant advancements in classification rates for these types. 16QAM and 32QAM can be correctly distinguished at the SNR. Other modulation types can achieve a favorable classification effect. As SNRs are elevated, the used network learns more hidden signal traits and accomplishes the desired classification results.

The t-distributed stochastic neighbor embedding (t-SNE) (Van der Maaten and Hinton, 2008) is used to analyze the features of underwater modulation signals in Figure 7. A total of 400,000 testing signal data randomly selected from dataset are adopted for the experiment. The output results are extracted as the obtained recognition features in the last dense layer. Most modulation styles can be separated from each other by the proposed method, which constructs the feature map for the effective identification. The previous classification results match the testing effects. There are

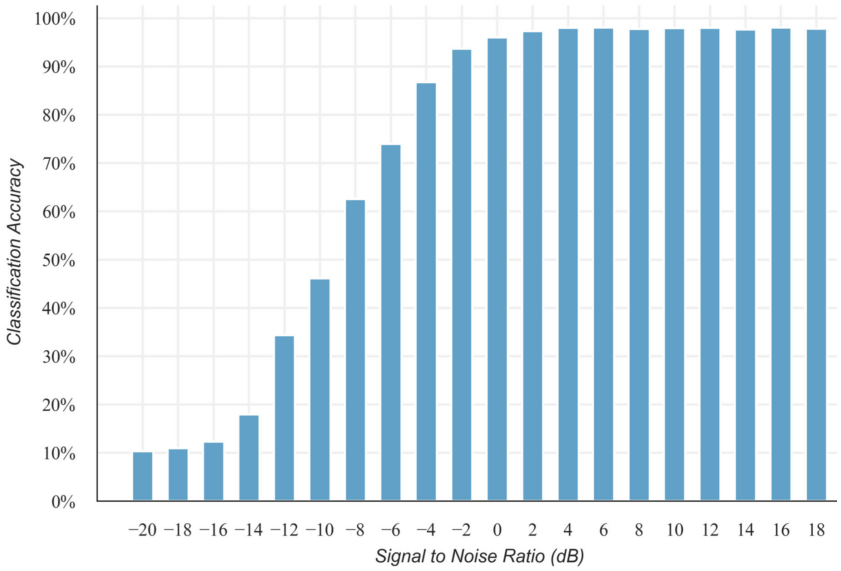

FIGURE 4
The classification results of the proposed method.

some overlaps between QPSK, 8PSK, 16QAM, and 64QAM for features. The main reason for the phenomenon is that they are polluted by the underwater emission attenuation, which have similar constellation distributions at low SNRs. The results illustrate that the proposed method exhibits better view of the class types in the underwater acoustic signal modulations.

In Figure 8, the proposed network (ProposedNet) is compared with ShuffleNet (Ma et al., 2018), MobileNet (Howard et al., 2019), Xception Chollet (2017), ANResNet (Liang et al., 2021) and IAFNet (Wang et al., 2022a). ShuffleNet is characterized by its wide-structured network design. MobileNet and Xception are lightweight neural networks that utilize stacking of smaller convolutional kernels for efficiency. Meanwhile, ANResNet and IAFNet represent networks with more complex structures in the field of underwater acoustic signal recognition. ShuffleNet has better classification results than the proposed network in the range of −20 dB and −12 dB, and the method has never really worked at a low SNR. MobileNet, ANResNet, and Xception have almost the same classification result as the proposed network. There is a similar result between IAFNet and the proposed network from −20 dB to

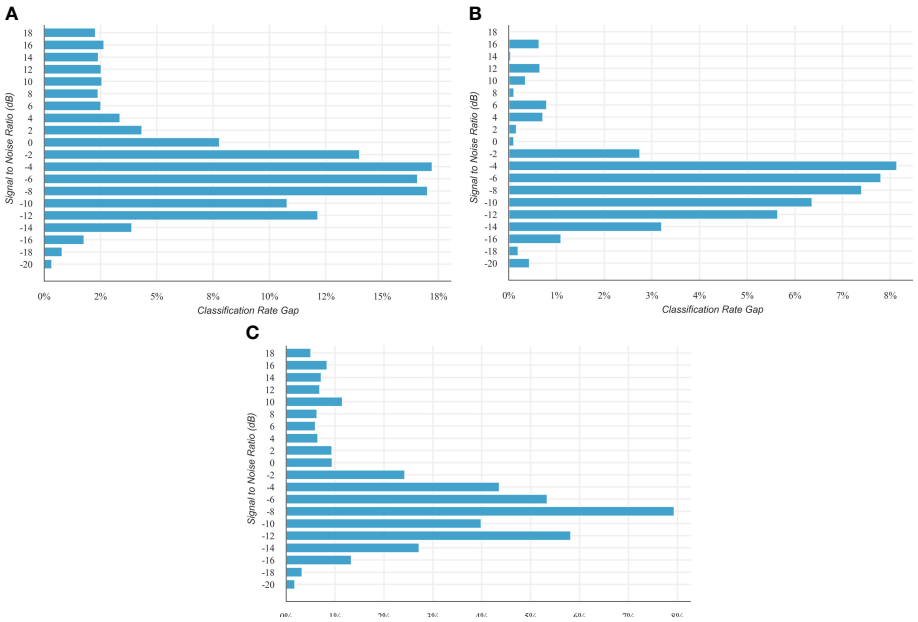

FIGURE 5
Modulation classification results for varying SNRs. **(A)** The classification rate of DWC is higher than that of CNN. **(B)** The classification rate of DWC is higher than that of SeparableCNN. **(C)** The classification rate of DWC is higher than that of TransposeCNN.

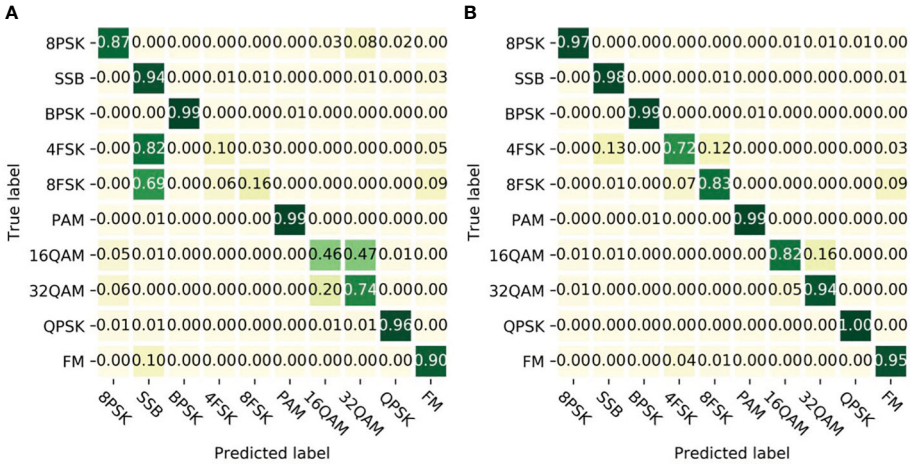

FIGURE 6
Modulation classification results for varying SNRs. **(A)** SNR = -6 dB. **(B)** SNR = -4 dB.

−16 dB, and the performance of IAFNet does not improve enough as SNRs strengthen. From −12 dB to −4 dB, the proposed network is substantially higher than other networks in the classification rate, which is approximately 3.2%, 7.1%, 8.0%, 4.6%, and 29.2% higher than ShuffleNet, MobileNet, Xception, IAFNet, and ANResNet on average, respectively. It shows that the proposed network has a superior structure and obtains more advanced classification information of signals. After SNR = −4 dB, there is an impressive improvement in the classification effect of all networks. The proposed network is superior to the other four networks. ShuffleNet, MobileNet, and Xception demonstrate similar trends in classification performance, particularly in key performance metrics such as accuracy and recall rate, showing comparable results when processing specific types of datasets. They are less effective than the proposed network, which is better by approximately 11.2% and 11.4% than ShuffleNet and MobileNet, respectively. Meanwhile, the proposed network outperforms Xception and ANResNet by approximately 12.4% and 8.7%, which is far superior to IAFNet. It is due to the network structure that enriches the trait extraction of signals, which performs better than the wide network structure of the lightweight network of ShuffleNet, MobileNet, and Xception, the commonly used underwater acoustic recognition network structure of ANResNet and IAFNet.

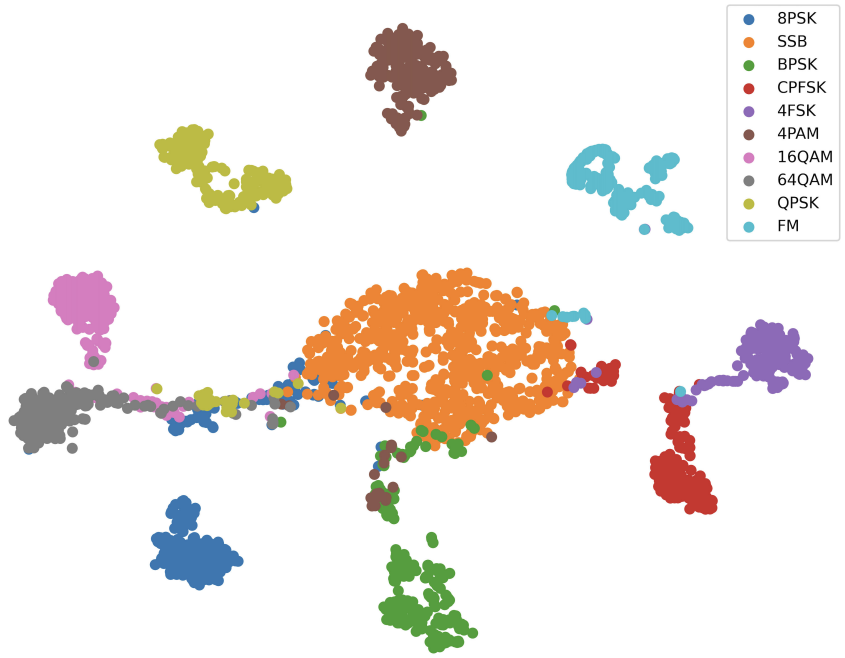

FIGURE 7
t-SNE visualization for testing features learned from the proposed method.

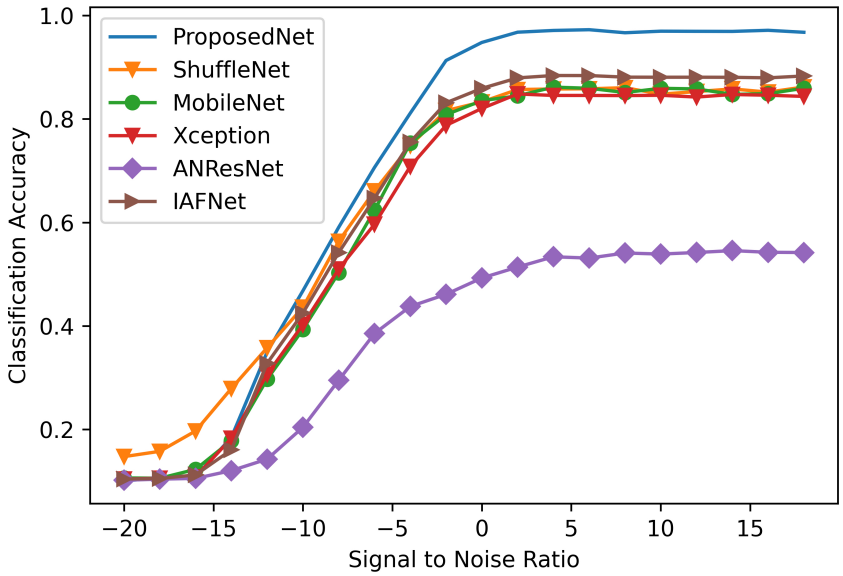

FIGURE 8
Classification results between different networks.

The proposed network compares the epoch time (training time) and the parameter size with other network models in Table 2, which were obtained on ubuntu 18.04, tensorflow version 2.12, CPU i7, and GPU 2080ti.

Compared to the proposed network, the parameter sizes of ShuffleNet, MobileNet, and ANResNet2 are four times, five times, and six times larger, respectively. ANResNet and IAFNet have a complex structural style, and have a fairly large parameter size. They are nearly 20 times and 52 times larger than the proposed network in terms of parameter size, respectively. The proposed network has the shortest amount of epoch time, which is approximately 1/3, 1/4, 1/5, 1/9, and 1/19 the epoch time of ShuffleNet, MobileNet, Xception, ANResNet, and IAFNet, respectively. The epoch time is related not only to the parameter size but also to the complexity of the network structure. The structural design of the used network is more efficient with the signal trait exchange of the multi-routing structure, which has a smaller parameter size and a shorter training time. The proposed network is more appropriate to apply and embed in a real underwater communication system.

TABLE 2 The parameter size and epoch time of different network models.

The network model	Epoch time (s)	Total parameters
The proposed network	6	868,762
ShuffleNet	19	3,166,848
MobileNet	24	3,718,570
Xception	29	5,334,562
ANResNet	56	20,772,894
IAFNet	112	44,961,468

4 Conclusion

In an underwater acoustic environment, this algorithm proposes the novel fusion network for AMC. The proposed method consists of both Transformer and the spatial conversion CNN module, and the results show the effectiveness in underwater modulation classification. The Transformer network can input signal sequences and has the attention mechanism, which can obtain more hidden distinguishing signal information at lower SNRs. The embedded DWC can acquire deep representations in spatial domain, which improves the recognition accuracy at lower SNRs. In future work, research should center on improving the classification accuracy in the less than −15 dB SNR range.

Nomenclature

Resource identification initiative

No specific resources requiring an RRID were used in this study.

Life science identifiers

This study did not involve the description of new species or nomenclatural acts registered with ZOOBANK; therefore, no Life Science Identifiers (LSIDs) are included.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding author.

Author contributions

YW: Writing – review & editing, Writing – original draft, Visualization, Methodology, Conceptualization. JX: Writing – review & editing, Project administration, Methodology, Investigation. XC: Writing – original draft, Validation, Software, Investigation. QW: Supervision, Formal Analysis, Writing – review & editing, Project administration. NT: Writing – review & editing, Resources, Project administration, Data curation.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was financially supported by Scientific Research Startup Foundation of Taishan University (No. Y-01-2020016), Shandong Provincial Natural Science Foundation (No. ZR2022MF347).

Acknowledgments

We wish to extend our sincere thanks to College of Electronic Engineering, Faculty of Information Science and Engineering, Ocean University of China, for their invaluable assistance with the development of the computational models. Their insights and expertise were pivotal to the successful completion of this project, and their contributions to our discussions and their technical input have been greatly appreciated. We are also grateful for the supportive environment provided by the Faculty of Information Science and Engineering at Ocean University of China, which has been instrumental in facilitating our research endeavors.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abu-Romoh, M., Aboutaleb, A., and Rezki, Z. (2018). Automatic modulation classification using moments and likelihood maximization. *IEEE Commun. Lett.* 22, 938–941. doi: 10.1109/LCOMM.2018.2806489

Cai, X., Xu, W., Wang, L., and Kaddoum, G. (2022). Joint energy and correlation detection assisted non-coherent ofdm-dcsk system for underwater acoustic communications. *IEEE Trans. Commun.* 70, 3742–3759. doi: 10.1109/TCOMM.2022.3169227

Chithaluru, P., Stephan, T., Kumar, M., and Nayyar, A. (2022). An enhanced energy-efficient fuzzy-based cognitive radio scheme for iot. *Neural Comput. Appl.* 34, 19193–19215. doi: 10.1007/s00521-022-07515-8

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions, in *Proceedings of the IEEE conference on computer vision and pattern recognition*. (Honolulu, Hawaii: IEEE), 1251–1258.

Dong, S., Wang, P., and Abbas, K. (2021). A survey on deep learning and its applications. *Comput. Sci. Rev.* 40, 100379. doi: 10.1016/j.cosrev.2021.100379

Dong, Y., Shen, X., Jiang, Z., and Wang, H. (2021). Recognition of imbalanced underwater acoustic datasets with exponentially weighted cross-entropy loss. *Appl. Acoustics* 174, 107740. doi: 10.1016/j.apacoust.2020.107740

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fang, T., Wang, Q., Zhang, L., and Liu, S. (2022). Modulation mode recognition method of noncooperative underwater acoustic communication signal based on spectral peak feature extraction and random forest. *Remote Sens.* 14, 1603. doi: 10.3390/rs14071603

Gao, D., Hua, W., Su, W., Xu, Z., and Chen, K. (2022). Supervised contrastive learning-based modulation classification of underwater acoustic communication. *Wireless Commun. Mobile Comput.* 2022, 1–10. doi: 10.1155/2022/3995331

Hamee, H. M., and Wadi, J. (2015). Automatic modulation recognition for mfsk using modified covariance method. *Int. J. Electrical Comput. Eng. (IJECE)* 5, 429–435. doi: 10.11591/ijece.v5i3

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., et al. (2019). Searching for mobilenetv3, in *Proceedings of the IEEE/CVF international conference on computer vision*. (Seoul, South Korea: IEEE) 1314–1324.

Hreshee, S. S. (2020). Automatic recognition of the digital modulation types using the artificial neural networks. *Int. J. Electrical Comput. Eng. (2088-8708)* 10:5871–5882. doi: 10.11591/ijece.v10i6

Hu, G., Wang, K., and Liu, L. (2021). Underwater acoustic target recognition based on depthwise separable convolution neural networks. *Sensors* 21, 1429. doi: 10.3390/s21041429

Huang, Z., Li, S., Yang, X., and Wang, J. (2022). Oae-eeknn: An accurate and efficient automatic modulation recognition method for underwater acoustic signals. *IEEE Signal Process. Lett.* 29, 518–522. doi: 10.1109/LSP.2022.3145329

Li, X., Dong, F., Zhang, S., and Guo, W. (2019). A survey on deep learning techniques in wireless signal recognition. *Wireless Commun. Mobile Comput.* 2019, 1–12. doi: 10.1155/2019/5629572

Liang, Z., Tao, M., Wang, L., Su, J., and Yang, X. (2021). Automatic modulation recognition based on adaptive attention mechanism and resnext wsl model. *IEEE Commun. Lett.* 25, 2953–2957. doi: 10.1109/LCOMM.2021.3093485

Liu, X., Xu, B., Wang, X., Zheng, K., Chi, K., and Tian, X. (2022). Impacts of sensing energy and data availability on throughput of energy harvesting cognitive radio networks. *IEEE Trans. Vehicular Technol.* 72, 747–759. doi: 10.1109/TVT.2022.3204310

Ma, N., Zhang, X., Zheng, H.-T., and Sun, J. (2018). "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)* (Munich, Germany: Springer), 116–131.

Menghani, G. (2023). Efficient deep learning: A survey on making deep learning models smaller, faster, and better. *ACM Comput. Surveys* 55, 1–37. doi: 10.1145/3578938

Song, X., Li, J., Cai, T., Yang, S., Yang, T., and Liu, C. (2022). A survey on deep learning based knowledge tracing. *Knowledge-Based Syst.* 258, 110036. doi: 10.1016/j.knosys.2022.110036

Teekaraman, Y., Manoharan, H., Basha, A. R., and Manoharan, A. (2023). Hybrid optimization algorithms for resource allocation in heterogeneous cognitive radio networks. *Neural Process. Lett.* 55, 3813–3826. doi: 10.1007/s11063-020-10255-2

Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-sne. *J. Mach. Learn. Res.* 9. 2579–2605.

Wang, H., Wang, B., and Li, Y. (2022a). Iafnet: Few-shot learning for modulation recognition in underwater impulsive noise. *IEEE Commun. Lett.* 26, 1047–1051. doi: 10.1109/LCOMM.2022.3151790

Wang, H., Wang, B., Wu, L., and Tang, Q. (2022b). Multihydrophone fusion network for modulation recognition. *Sensors* 22, 3214. doi: 10.3390/s22093214

Wang, Y., Zhang, H., Sang, Z., Xu, L., and Gulliver, T. A. (2019). Modulation classification of underwater communication with deep learning network. *Comput. Intell. Neurosci.* 2019, 1–12. doi: 10.1155/2019/9142753

Xu, M., Yoon, S., Fuentes, A., and Park, D. S. (2023). A comprehensive survey of image augmentation techniques for deep learning. *Pattern Recognit.* 137, 109347. doi: 10.1016/j.patcog.2023.109347

Zhai, Y., Li, J., Feng, H., and Hong, F. (2023). Application research of polar coded ofdm underwater acoustic communications. *EURASIP J. Wireless Commun. Netw.* 2023, 26. doi: 10.1186/s13638-023-02236-5

Zhang, Y., Li, C., Wang, H., Wang, J., Yang, F., and Meriaudeau, F. (2022). Deep learning aided ofdm receiver for underwater acoustic communications. *Appl. Acoustics* 187, 108515. doi: 10.1016/j.apacoust.2021.108515

Zhang, W., Tait, A., Huang, C., Ferreira de Lima, T., Bilodeau, S., Blow, E. C., et al. (2023). Broadband physical layer cognitive radio with an integrated photonic processor for blind source separation. *Nat. Commun.* 14, 1107. doi: 10.1038/s41467-023-36814-4

Zhang, W., Yang, X., Leng, C., Wang, J., and Mao, S. (2022). Modulation recognition of underwater acoustic signals using deep hybrid neural networks. *IEEE Trans. Wireless Commun.* 21, 5977–5988. doi: 10.1109/TWC.2022.3144608

Zheng, T., Jing, L., Long, C., He, C., and Yin, H. (2023). Frequency domain direct adaptive turbo equalization based on block normalized minimum-ser for underwater acoustic communications. *Appl. Acoustics* 205, 109266. doi: 10.1016/j.apacoust.2023.109266

Zheng, C., Wu, W., Chen, C., Yang, T., Zhu, S., Shen, J., et al. (2023). Deep learning-based human pose estimation: A survey. *ACM Comput. Surveys* 56, 1–37. doi: 10.1145/3603618