Union-Domain Knowledge Distillation for Underwater Acoustic Target Recognition

Xiaohui Chu, Haoran Duan, Zhenyu Wen, Lijun Xu, Runze Hu, Wei Xiang, *Senior Member, IEEE*

***Abstract*—Underwater Acoustic Target Recognition (UATR) can be significantly empowered by advancements in deep learning (DL). However, the effectiveness of DL-based UATR methods is often constrained by the limited computing resources available on underwater platforms. Most of the existing knowledge distillation (KD) strategies try to build lightweight DL models, but these strategies rarely consider the acoustic properties of underwater environments, making them less efficient for UATR tasks. Thus, fully harnessing the potential of DL techniques while ensuring the model's practicality, is one of the urgent problems to be solved in UATR research. In this work, we introduce the Union-Domain Knowledge Distillation (UDKD) to establish an accurate and lightweight UATR model. UDKD integrates two KD strategies: Dual-frequency Band Distillation (DBD) and Cross-domain Masked Distillation (CMD). DBD improves the learning process for a simple student model by decoupling the knowledge of spectrograms into the local structural (i.e., line spectra) and global composition (i.e., propagation patterns) aspects. CMD reduces redundant information from the Fourier Transform process, enabling the student model to concentrate on essential signal elements and to learn underlying time-frequency distribution. Extensive experiments on two real-world oceanic datasets confirm the superior performance of UDKD compared to existing KD methods, i.e., achieving an accuracy of 94.81% (↑ 3.19% v.s. 91.62%). Notably, UDKD showcases a 10.5% improvement in the prediction accuracy of the lightweight student model.**

***Index Terms*—Acoustic recognition, computer vision, knowledge distillation, model compression, frequency domain.**

I. Introduction

UNDERWATER Acoustic Target Recognition (UATR) [1]–[4] classifies the target types based on received acoustic signals, which is crucial for understanding the marine environment. The application of UATR is broad, spanning from the exploration of oceanic resources to the development of advanced marine apparatus [5]–[7]. However, unlike conventional speech recognition technologies, UATR technology faces significant engineering challenges, particularly the energy scarcity in underwater platforms. This leads to very limited computational resources, and thus becoming problematic given the high computational demands of deep learning (DL) techniques. To mitigate these constraints, there is a growing interest in model light-weighting strategies, such as quantization [8], knowledge distillation [9], and neural architecture search [10], [11]. Among these, knowledge distillation (KD) emerges as a popular approach. It transfers the performance of a complex, high-parameter deep model (the teacher) to a simple, low-parameter light-weight model (the student). This offers more flexibility and robustness for the UATR tasks.

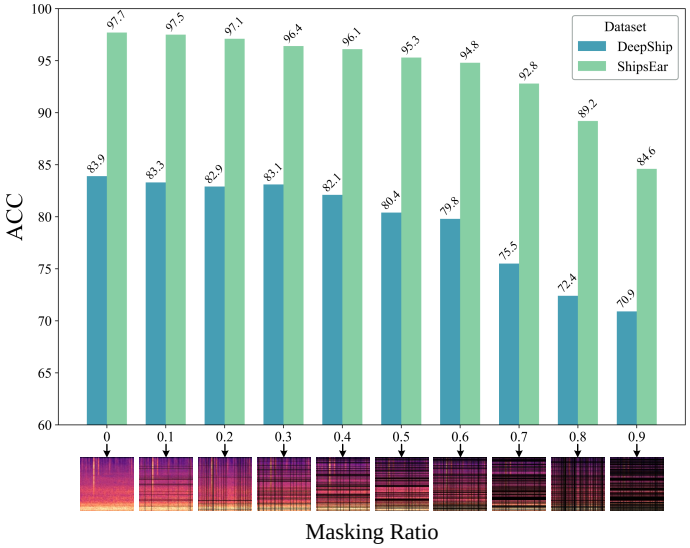


Fig. 1. The recognition accuracy (ACC) of teacher model under different masking ratios on DeepShip and ShipsEar datasets. The accuracy does not decrease significantly as the masking area on the input spectrograms increases, indicating that a great deal of redundant information exists in the spectrograms.

Knowledge distillation is categorized mainly into logit-based [12]–[14] and feature-based methods [15]–[17]. Logit-based KD involves teaching the student model to emulate the teacher's output logits, while feature-based KD allows the student to learn from the teacher's intermediate feature maps. Feature-based KD typically offers superior performance [18] by imparting more comprehensive knowledge, making it an appealing choice for UATR. However, the direct application of existing feature-based KD methods to the UATR task is often inefficient due to the inherent differences between traditional digital imagery and acoustic data represented in spectrograms.

Specifically, UATR models commonly utilize spectrograms, such as Mel spectrograms, to visualize changes in frequency content over time. In contrast to digital images that display rich semantic content through a blend of high-level and low-level features [19], Mel spectrograms lack explicit content-related information. Therefore, it necessitates more effective inter-

X. Chu, R. Hu, and L. Xu are with the School of Information and Electronics, Beijing Institute of Technology, Beijing, China (e-mail: 3120225380@bit.edu.cn, hrzlpk2015@gmail.com, xlj@bit.edu.cn).

H. Duan is with the Department of Computer Science, Durham University, Durham, U.K. (e-mail: haoran.duan@ieee.org).

Z. Wen is with the School of Computing Science, Zhejiang University of Technology, Hangzhou, China (e-mail: wenluke427@gmail.com).

W. Wei is School of Computing Engineering and Mathematical Sciences, La Trobe University, Melbourne 3086, Australia (e-mail: w.xiang@latrobe.edu.au).

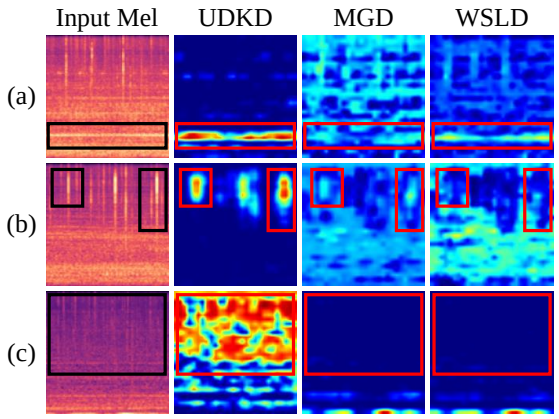


Fig. 2. Visualization of feature heatmaps from student model using different KD frameworks, i.e., our proposed UDKD, MGD [20], and WSLD [21]. The student from UDKD can better grasp the essential features such as the distinct line spectral and underlying time-frequency distribution pattern.

pretation mechanisms in the KD process. Furthermore, while digital images exhibit significant pixel contrast variations, changes in Mel spectrograms are comparatively subtle, leading to information redundancy. For instance, as demonstrated in Fig. 1, increasing the masking ratio in spectrograms does not significantly impact the model accuracy. Such an information redundancy in spectrograms restrains the efficiency of traditional KD approaches.

The discrepancies between the acoustical spectrograms and the digital images prompt two rationales for designing a KD paradigm specific for the UATR task: ***(1) Knowledge learning in the frequency domain.*** Establishing the feature representation in the frequency domain allows for richer patterns, which provide diverse interpretations for original acoustical data [22]. ***(2) Masked Image Modeling (MIM).*** MIM [23]–[25] scheme masks a significant amount of image information to eliminate the redundancy, which can be potentially beneficial for dealing with the redundant information of Mel spectrograms.

To this end, we propose a new knowledge distillation framework, namely Union-Domain Knowledge Distillation (UDKD), which involves two strategies for effective knowledge transfer in acoustical spectrograms. First, we present the Dual-frequency Band Distillation (DBD) module that engages with the frequency domain characteristics of spectrograms. DBD performs the Fourier Transform to convert feature vectors into their frequency domain representations [22]. Essential frequencies are then isolated to allow for the precise transfer of the teacher's high and low-frequency features to the student. Consequently, the student model can focus on crucial spectral features and pivotal time periods (high-frequency), i.e., Fig. 2 (a) and (b), while concurrently assimilating the global propagation patterns of underwater acoustic signals (low-frequency).

Second, to mitigate the information redundancy in spectrograms, we propose the Cross-domain Masked Distillation (CMD) module. Traditional random patch-masked strategies [26], while effective to a degree, often compromise the structural integrity of the acoustic data. In comparison, CMD respects time-frequency regularities by masking all frequency domain information within a specific time slot and obfuscating all time domain knowledge for a designated frequency band. This masking strategy naturally simulates the interference of the marine environment, such as transient noises or silent intervals. Consequently, CMD enhances the student's ability to understand the complex interactions between frequency bands and time segments, and thus grasping the underlying time-frequency distribution patterns of acoustic signals (see Fig. 2(c)). A supervision task is further introduced into the CMD to improve the student's efficiency in learning from the masked areas. With these two modules, UDKD significantly enhances the student's learning capacity. To summarize, this paper makes the following main contributions:

- We introduce a novel knowledge distillation framework, tailored for UATR tasks, called Union-Domain Knowledge Distillation (UDKD). UDKD facilitates knowledge transfer across both time and frequency domains simultaneously, which represents the first of its kind in the UATR task.
- We develop the Dual-frequency Band Distillation (DBD) module to analyze spectrograms from a frequency perspective. DBD enables the student to learn key features, i.e., line spectral features (high-frequency) and propagation patterns of acoustic signals (low-frequency), establishing a more holistic mechanism for knowledge distillation.
- To address the issue of information redundancy in the spectrograms, we design the Cross-domain Masking Distillation (CMD) module. It establishes a masked autoencoder paradigm that includes an effective masking strategy and a sophisticated decoder module. CMD facilitates efficient learning for the complex time-frequency distribution patterns in acoustic signals.
- On two real-world oceanic datasets, our proposed UDKD further enhances the effectiveness of feature-based distillation, achieving the recognition accuracy of 94.81% for a simple student (3-layer convolutional model), which is 10.5% higher than the baseline, and also outperforms the existing SOTA methods.

II. Related Work

A. Underwater Acoustic Target Recognition

Due to the complex acoustic properties in the marine environment, UATR tasks face great challenges, such as ambient noise effects, propagation losses, and multi-path interference. Traditional recognition methods are affected by physical and psychological factors, which may not generalize well in practical applications. In the last decades, using machine learning and deep learning to identify underwater acoustic signals has gained a great deal of attention.

The research of UATR based on machine learning mainly focuses on feature extraction of acoustic signals and classifier design. The main feature extraction methods include the Hilbert-Huang Transform (HHT) [27], the Short-Time Fourier Transform (STFT) [28], the Mel Frequency Cepstrum Coefficient (MFCC) [29]. In addition, the classifiers represented by Decision Tree (DT) [30], Support Vector Machine (SVM) [31] and other machine learning methods were applied to

UATR. These methods significantly improved the performance of UATR tasks. Furthermore, with the wide applications of deep neural networks (DNN) [32], researchers have begun to apply DNN to solve the UATR problems, and remarkable progress has been made. For instance, Liu et al. [3] designed a recognition method based on a convolutional recurrent neural network. Since the Mel spectrograms can only represent static features, the dynamic features in delta spectrograms were introduced to construct 3-D Mel spectrogram input. Tian et al. [2] explored a deep convolutional stack network suitable for perceiving underwater radiation noise. On this basis, a multi-scale residual module was presented to classify underwater acoustic signals. Xie et al. [4] established a contrastive learning paradigm that included audio, spectrogram, and text modes, integrating relevant information from different perspectives. Zhou et al. [1] adopted a joint training framework for noise-robust underwater acoustic recognition, in which, a cross-attention mechanism was proposed to simulate the noisy environment.

B. Knowledge Distillation

Knowledge Distillation (KD) is a model compression and acceleration technique. During the KD process, the student can learn the diverse knowledge and generalization capabilities of the teacher, resulting in a smaller computational complex and faster running speed while maintaining high accuracy. The logit-based distillation was first proposed by Hinton et al. [9], which used the soft logit outputs of the teacher model as the target labels to train the student model. Unlike conventional one-hot labels, which cannot fully describe the nuanced relationships between samples and their corresponding labels due to their overconfidence, these soft logit outputs assign higher probabilities to classes that are similar to the correct class. They consider that other classes information can provide some additional knowledge, i.e., "dark knowledge", representing instance-to-class and class-to-class similarity information. On this basis, several improvement methods were proposed to further enhance the efficiency of knowledge transferring. CCKD [33] took correlation between instances as the transferred knowledge, thus mimicking the characteristics of intra-class samples aggregation and inter-class samples separation in the teacher's feature space. SSKD [34] introduced self-supervised learning to help student extract more comprehensive knowledge from teacher. KDExplainer [35] analyzed the effects of soft labels in training the student, and further designed virtual attention module to coordinate different types of knowledge conflicts. DKD [36] divided the logit knowledge into target knowledge and non-target knowledge to explore the richer semantic information in the deeper layers of the model.

The feature-based distillation was originally proposed in [37], which forced the student model to mimic the intermediate feature maps of the teacher model. This paradigm broadened the definition of knowledge, i.e., knowledge was not only a response to the outputs of a larger model, but also implicit in some intermediate layer representations. Since then, various methods have been proposed to facilitate knowledge transfer by changing matching methods or replacing matching features. AT [17] enabled student to mimic the attention maps of a strong teacher to significantly improve its performance. SP [38] calculated the pair-wise similarities between teacher and student, so that the teacher and student generated similar activation for the same samples. RKD [14] considered the mutual relations of data instances as knowledge that was transferred between teacher and student. reviewKD [39] proposed a new knowledge transfer mechanism that used multi-level information of teacher to guide single-level learning of student. OFD [40] studied various aspects of knowledge distillation, making distillation losses synergistic among teacher/student feature transformation and the feature position.

III. Methodology

In this section, we first provide the preliminaries regarding the principles of Mel spectrograms. Then, we elaborate on the proposed Union-Domain Knowledge Distillation framework, dubbed UDKD, which involves two kinds of knowledge distillation: Dual-frequency Band Distillation (DBD) and Cross-domain Masked Distillation (CMD).

A. Principles of Mel spectrograms

The acoustic signals can be transformed into various formats of spectrograms, which are generally treated as input data to train the UATR model. In Fig. 3, we present four commonly used spectrograms in the UATR tasks, including Short-Time Fourier Transform (STFT), Mel Frequency Cepstrum Coefficient (MFCC), Constant Q Transform (CQT) [41], and Mel spectrograms. As can be observed, the STFT and CQT exhibit comprehensive time-frequency information of the targets. However, the spectral characteristics and crucial time periods of the targets are not clearly highlighted. As for the MFCC, the discrete cosine transform filters out considerable time-frequency characteristics of the data. In comparison, Mel spectrograms earn three advantages: (1) Mel spectrograms preserve the spectral features more complete (see the black box in Fig. 3). (2) Mel spectrograms avoid blocky representations with low resolution (such as STFT) and point-like noise (such as CQT), as shown in azure box in Fig. 3. (3) Mel spectrograms exhibit the clearer time-frequency variation patterns. Therefore, this work utilizes the Mel spectrograms as the inputs for model training.

In the process of generating the Mel spectrogram, the acoustic signal is converted to the Mel scale in the frequency domain by applying a series of triangular band-pass filters. The Mel scale mimics the human's auditory perception towards different frequencies of sounds, especially the high sensitivity in the low-frequency range and the low sensitivity in the high-frequency interval. The generation of Mel spectrograms is summarized in Fig. 4, which involves five stages.

1) The acoustic signal is pre-processed through framing and windowing. Due to the short-time stationary characteristics of the signals, it is necessary to perform framing. The purpose of windowing is to mitigate spectrum leakage.

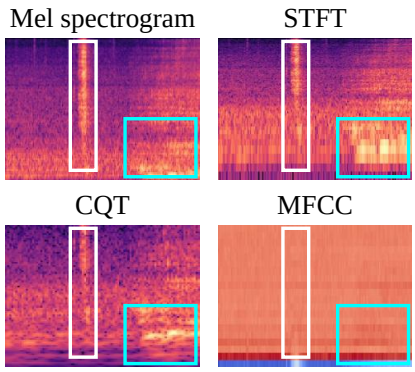


Fig. 3. Visual representations of different spectrograms. Mel spectrograms preserve rich visual information.

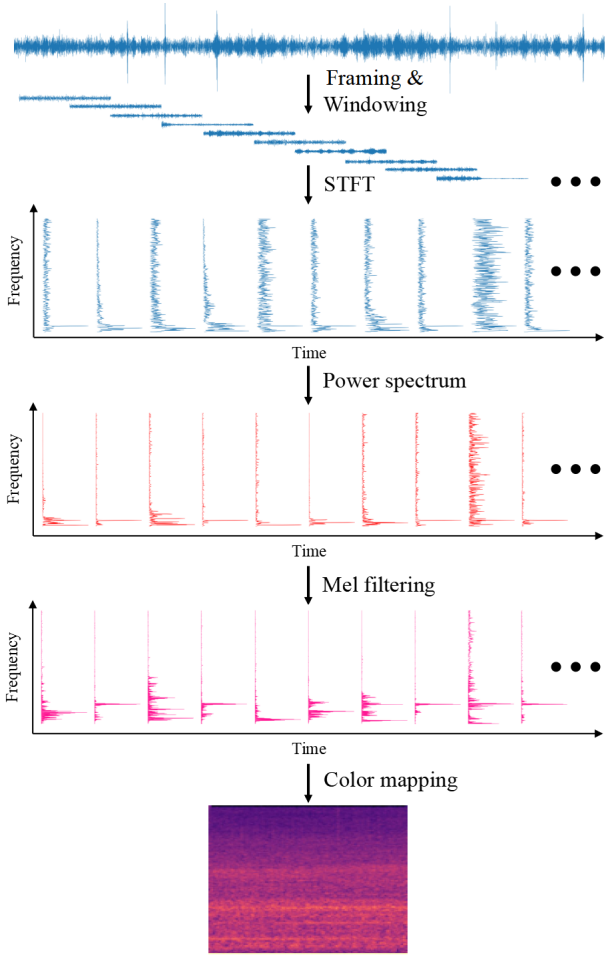


Fig. 4. Generation process of Mel spectrograms.

2) STFT is utilized to transform the signal from the time domain into the time-frequency domain. The transformed signals are squared to obtain the linear power spectrum on the Hz frequency scale:

$$\boldsymbol{S}(u,v) = |\,\mathrm{STFT}(\boldsymbol{y}(n))|^2, \tag{1}$$

where $\boldsymbol{y} \in \mathbb{R}^{N\times 1}$ denotes the signal and N represents the number of sampling points per frame of the signal. $\boldsymbol{S} \in \mathbb{R}^{U\times V}$ indicates the linear power spectrum for the v-th time frame and the u-th frequency bin, where $v \in 1 \sim V$ and $u \in 1 \sim U$.

3) A set of Mel filter banks $\boldsymbol{B}$ are applied to convert the linear power spectrum into the non-linear power spectrum on the Mel frequency scale, written by:

$$\boldsymbol{X}(e,v) = \mathrm{EIN}(\boldsymbol{S},\boldsymbol{B}), \tag{2}$$

where $\boldsymbol{X} \in \mathbb{R}^{E\times V}$ denotes the Mel spectrogram, E represents the number of Mel bands, and $\mathrm{EIN}(\cdot,\cdot)$ indicates the Einstein summation convention [42].

4) A logarithmic operation is performed to compress the non-linear power spectrum obtained in Stage 3) and map it onto the frequency bands.

5) Finally, we save the mapped Mel spectrogram in the form of a color image, formulated as $\boldsymbol{X} \in \mathbb{R}^{C_M\times H_M\times W_M}$, where C_M, H_M, and W_M represent the channel number, height, and width of the Mel spectrogram, respectively.

B. Overall Framework

The architecture of Union-Domain Knowledge Distillation (UDKD) is shown in Fig. 5. Let $\boldsymbol{F}_{S,i} \in \mathbb{R}^{C_i\times H_i\times W_i}$ and $\boldsymbol{F}_{T,i} \in \mathbb{R}^{C_i\times H_i\times W_i}$ be the intermediate feature maps from the i-th layer of the student model S and teacher model T, respectively. We first utilize a 1×1 convolutional layer on $\boldsymbol{F}_{S,i}$ to align the channel dimension between the $\boldsymbol{F}_{S,i}$ and the $\boldsymbol{F}_{T,i}$. Then, $\boldsymbol{F}_{S,i}$ and $\boldsymbol{F}_{T,i}$ are sent to Dual-frequency Band Distillation (DBD) to force student to mimic the teacher's frequency domain features. DBD adopts the pre-defined filters to obtain the low-frequency and high-frequency feature maps of the student and teacher, and then performs the standard feature-based distillation between the corresponding features. Finally, a Cross-domain Masked Distillation (CMD) is introduced to encourage students to learn the time-frequency distribution patterns of acoustic signals. We mask $\boldsymbol{F}_{S,i}$ using a dedicatedly designed masking strategy while maintaining the whole input for $\boldsymbol{F}_{T,i}$, and the masked feature map is then sent into a decoder to reconstruct the latent representation. We calculate the distillation loss between the reconstructed student's feature maps and the complete teacher's feature maps. The above process can be formulated as:

$$\begin{aligned}\mathcal{P} = \sum_{i=1}^{K_S}[&\mathcal{P}_{\mathrm{fet}}\left(f_{\mathrm{DBD}}\left(\boldsymbol{F}_{T,i}\right), f_{\mathrm{DBD}}\left(\varphi\left(\boldsymbol{F}_{S,i}\right)\right)\right) + \\ &\mathcal{P}_{\mathrm{fet}}\left(\boldsymbol{F}_{T,i}, f_{\mathrm{CMD}}\left(\varphi\left(\boldsymbol{F}_{S,i}\right)\right)\right)],\end{aligned} \tag{3}$$

where K_S indicates the number of layers in the student model, $\mathcal{P}_{\mathrm{fet}}$ denotes the standard feature-based distillation paradigm, and φ represents the 1×1 convolutional layer. f_{DBD} and f_{CMD} denote the proposed DBD and CMD modules, respectively.

To ease the understanding of the proposed UDKD process, we utilize a standard convolutional neural network (CNN) as an example of the student model, which consists of three convolutional blocks (Conv-BN-ReLU), as shown in Fig. 6(a). For the teacher model, we choose the pre-trained ResNet-18, which is widely utilized as a backbone network across a diverse range of computer vision tasks [43], [44], as shown

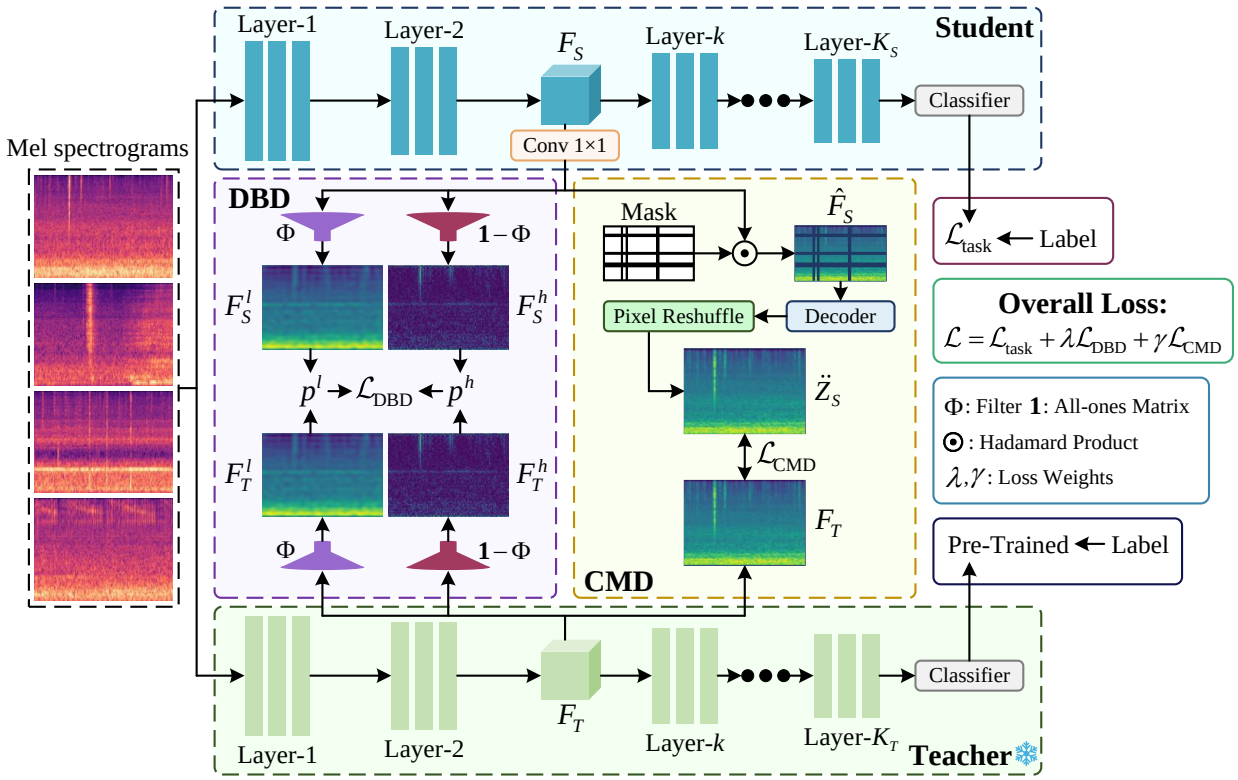
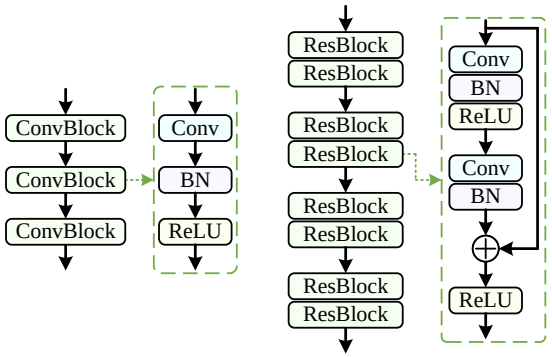


Fig. 5. Overall framework of the proposed UDKD.



(a) Student Model (b) Teacher Model

Fig. 6. The architecture of the exemplified student model and teacher model. In practice, the student and teacher can be arbitrary, i.e., CNN-based or vision transformer-based.

in Fig. 6(b). However, the architecture of the student model and teacher model can be arbitrary based on the application scenarios.

Our training pipeline can be divided into two steps. Firstly, we pre-train the teacher model on the DeepShip [45] and ShipsEar [46] datasets to obtain a robust teacher network. Secondly, we perform the overall distillation process, where the weights of the teacher model are frozen and the student model is trainable. Unlike digital images, which contain rich semantic content, Mel spectrograms primarily represent time-frequency distributions and lack high-level semantic details. Thus, we discard the final layer of the teacher model, as it is mainly responsible for capturing the abstract semantic features. Finally, we transfer knowledge from the first three layers of the teacher model to the corresponding layer of the student model in a one-to-one manner.

C. Dual-frequency Band Distillation

The Mel spectrogram is fundamentally a result of the Short-Time Fourier Transform. It converts the signal from the time domain to the time-frequency domain, and finally maps it into the form of a digital image. However, unlike conventional digital images, it is difficult to learn an effective classification pattern from the spatial domain solely in the Mel spectrogram. To this end, this paper shifts the focus to the frequency domain to exploit the underlying patterns from the Mel spectrograms. Specifically, the low-frequency components contain the propagation patterns of acoustic signals, while the high-frequency counterparts exhibit important line spectra and prominent time periods. These properties in the frequency domain are effectively complementary to those from the spatial domain. Therefore, we propose the Dual-frequency Band Distillation (DBD) module to transfer the low/high-frequency features of the intermediate feature maps between the student and the teacher.

The architecture of DBD is illustrated in Fig. 7, where the inputs to DBD are intermediate feature maps of student and teacher models. The channel dimension of the student's feature map is aligned with the teacher through a 1×1 convolutional layer. Given a feature map $\boldsymbol{F} \in \mathbb{R}^{C\times H\times W}$ (for simplification, we omit the subscript S, T, and i), we first

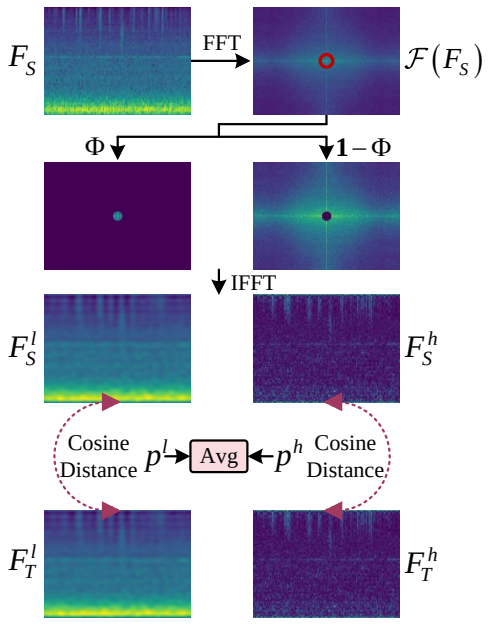


Fig. 7. Illustration of the Dual-frequency Band Distillation.

perform Discrete Fourier Transform (DFT) $\mathcal{F}$ to obtain the corresponding frequency representation:

$$\mathcal{F}(\boldsymbol{F})(a,b) = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} \boldsymbol{F}(h,w) e^{-j2\pi\left(\frac{ah}{H}+\frac{bw}{W}\right)}, \tag{4}$$

where $\mathcal{F}(\boldsymbol{F})(a,b)$ is the complex value at the coordinate of (a,b) on the frequency spectrum. e and j are Euler's number and the imaginary unit, respectively. Note that $\mathcal{F}$ operates on each channel independently.

We then design a filter $\boldsymbol{\Phi} \in \{0,1\}^{H\times W}$ with a circular shape to execute low/high-pass filtering. The value of $\boldsymbol{\Phi}$ is determined by an indicator function in Eq. (5), which separates the low frequencies and high frequencies based on the radius r of the circular filter:

$$\boldsymbol{\Phi}(a,b) = \begin{cases} 1, & \text{if } D\left((a,b),(O_H,O_W)\right) < r, \\ 0, & \text{otherwise}, \end{cases} \tag{5}$$

where (O_H, O_W) denotes the center of the feature map and $D(\cdot,\cdot)$ represents the Euclidean Distance. The filter shape is intuitively designed based on that the low-frequency information is aggregated in the center of the feature map. In addition, the circular filter removes the same amount of frequencies in all directions of the spectra, ensuring the balance between the low-frequency features and high-frequency features. We have discussed the effects of different filter shapes on the KD performance in Table V in Section IV.

With $\boldsymbol{\Phi}$, we can obtain the low-frequency feature map $\boldsymbol{F}^l$ and high-frequency feature map $\boldsymbol{F}^h$ as:

$$\begin{aligned} \boldsymbol{F}^l &= \mathcal{F}^{-1}(\mathcal{F}(\boldsymbol{F}) \odot \boldsymbol{\Phi}), \\ \boldsymbol{F}^h &= \mathcal{F}^{-1}(\mathcal{F}(\boldsymbol{F}) \odot (\mathbf{1}-\boldsymbol{\Phi})), \end{aligned} \tag{6}$$

where $\mathcal{F}^{-1}$ denotes the inverse Fourier transform, $\odot$ indicates the Hadamard product, and $\mathbf{1}$ is an all-ones matrix. In practice, $\mathcal{F}$ and $\mathcal{F}^{-1}$ can be calculated using Fast Fourier Transform (FFT) algorithm.

Finally, we distill $\boldsymbol{F}^l$ and $\boldsymbol{F}^h$ from teacher to student by:

$$\boldsymbol{p}_i^\dagger = 1 - \left\langle \boldsymbol{F}_{S,i}^\dagger, \boldsymbol{F}_{T,i}^\dagger \right\rangle = 1 - \frac{\boldsymbol{F}_{S,i}^\dagger}{\left\|\boldsymbol{F}_{S,i}^\dagger\right\|_2} \cdot \frac{\boldsymbol{F}_{T,i}^\dagger}{\left\|\boldsymbol{F}_{T,i}^\dagger\right\|_2}, \dagger = \{l,h\}, \tag{7}$$

where $\boldsymbol{p}_i^\dagger$ is the cosine distance between low-frequency feature maps or high-frequency feature maps from the i-th layer of the student and teacher models, $\langle\cdot,\cdot\rangle$ denotes the cosine similarity, and $\|\cdot\|_2$ indicates L2-norm. The distillation loss for DBD is calculated by averaging $\boldsymbol{p}_i^l$ and $\boldsymbol{p}_i^h$, and summing them across layers in the model, written by:

$$\mathcal{L}_{\text{DBD}} = \sum_{i=1}^{K_S} \frac{\left(\boldsymbol{p}_i^l + \boldsymbol{p}_i^h\right)}{2}. \tag{8}$$

The proposed DBD utilizes the dedicatedly designed filters to decouple the frequency knowledge. This allows the model to focus on crucial frequency information, while avoiding a significant amount of duplicate information to confuse the student model.

D. Cross-domain Masked Distillation

Having improved the distillation effectiveness from the perspective of the frequency analysis, we further explore the time-frequency regularities exhibited in the Mel spectrograms. As illustrated in Fig. 1, the spectrograms have the nature of information redundancy, thus impeding the learning efficiency of the student. To address this, we designed the Cross-domain Masked Distillation (CMD) module to make the student learn the time-frequency regularities efficiently. CMD randomly masks the frequency bands and time periods in the student's feature maps while maintaining the whole feature maps for the teacher. We introduce a decoder-based supervision task to force the student to comprehend the intrinsic relationships between the masked regions and their surrounding area, thus reconstructing the disrupted time-frequency patterns. Ultimately, the student can learn the underlying time-frequency distribution patterns of acoustic signals.

1) The Masking Strategy:

The architecture of CMD is illustrated in Fig. 8. Specifically, we send all the intermediate feature maps of the student model into CMD. Given a feature map $\boldsymbol{F}_{S,i} \in \mathbb{R}^{C_i\times H_i\times W_i}$, we first adjust the channel number using a 1×1 convolutional layer to align it with the channel number of the teacher's feature map. We then align all the feature maps to the same spatial resolution through a convolutional layer with the kernel size s. The aligned feature $\tilde{\boldsymbol{F}}_{S,i}$ can be obtained as:

$$\tilde{\boldsymbol{F}}_{S,i} = \text{Conv}_{s\times s}\left(\text{Conv}_{1\times 1}\left(\boldsymbol{F}_{S,i}\right)\right), \tag{9}$$

where $\tilde{\boldsymbol{F}}_{S,i} \in \mathbb{R}^{C_{S,i}\times\tilde{H}\times\tilde{W}}$, $\tilde{H} = H_M/16$, $\tilde{W} = W_M/16$, and $s = H_i/\tilde{H} = W_i/\tilde{W}$. $C_{S,i}$ indicates the channel number of student's feature map, which is consistent with the channel number in teacher's feature map.

When interpreting Mel spectrograms from an acoustic perspective, they display higher energy (loudness) in certain frequency bands and specific time segments, reflecting how sounds vary in loudness and pitch over time. Correspondingly,

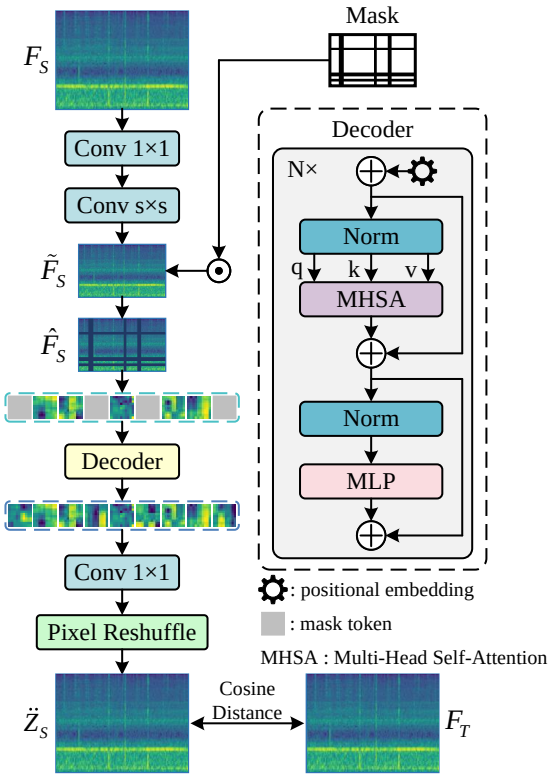


Fig. 8. Illustration of the Cross-domain Masked Distillation.

we propose masking all the frequency information according to the designated time slots, and occluding all the time knowledge at the selected frequency positions. Specifically, we first initialize an all-ones matrix and then randomly choose some time and frequency points based on the specified masking ratio. According to these selected time and frequency nodes, all the corresponding frequency and time information are masked, i.e., their values are set to 0. Finally, we generate a binary mask $\boldsymbol{M}$. The overall masking procedure with pseudo-code is presented in Algorithm 1.

Furthermore, the high complexity of the marine environment may cause acoustic signals to lose information in certain frequency bands and time intervals. The designed masking strategy can simulate such a situation, thus increasing the student's tolerance for the absence of acoustic information. This further enhances the practicality and robustness of the student. On the basis of $\boldsymbol{M}$, the masked feature $\hat{F}_{S,i}$ can be obtained by Hadamard product:

$$\hat{\boldsymbol{F}}_{S,i} = \tilde{\boldsymbol{F}}_{S,i} \odot \boldsymbol{M}. \tag{10}$$

2) The Decoder for Feature Recovery:

We leverage a decoder module to recover the missing information. Following the standard MAE [26] architecture, we first perform a series of dimensional transformations and information fusion on $\hat{\boldsymbol{F}}_{S,i}$, including: (1) Transform the shape of $\hat{\boldsymbol{F}}_{S,i}$ to $\mathbb{R}^{\tilde{H}\tilde{W}\times C_{S,i}}$ along the channel dimension. (2) Adjust the channel number of $\hat{\boldsymbol{F}}_{S,i}$ to C_D through a linear projection layer. (3) Add positional embedding $\boldsymbol{E}_{\text{pos}}$ to all tokens in the

Algorithm 1 Pseudo-code of masking procedure in a Python code style.

```
# F̃_{S,i}: the aligned feature, B × C_{S,i} × H̃ × W̃, B denotes the batch size.
# mask_ratio: the pre-defined masking ratio. The time and frequency points are randomly selected based on this parameter.
def MaskCreation(F̃_{S,i}, mask_ratio):
    B, _, H̃, W̃ = F̃_{S,i}.shape
    # initialize mask.
    mask = torch.ones([B, 1, H̃, W̃])
    for b in range(B):
        # calculate the time and frequency masking ratio.
        f_r = random.uniform(0, mask_ratio)
        t_r = mask_ratio - f_r
        # randomly select the time and frequency points.
        f_p = f_r ×H̃
        t_p = t_r ×W̃
        s_f_p = random.sample(list(range(H̃)), f_p)
        s_t_p = random.sample(list(range(W̃)), t_p)
        # set the values for the corresponding time and frequency positions to 0.
        mask[b, :, s_f_p, :] = 0
        mask[b, :, :, s_t_p] = 0
    return mask
```

$\hat{\boldsymbol{F}}_{S,i}$. These operations can be formulated as in:

$$\ddot{\boldsymbol{F}}_{S,i} = \text{Linear}\left(\text{Transform}\left(\hat{\boldsymbol{F}}_{S,i}\right)\right) + \boldsymbol{E}_{\text{pos}}, \tag{11}$$

where $\ddot{\boldsymbol{F}}_{S,i} \in \mathbb{R}^{\tilde{H}\tilde{W}\times C_D}$ and $\boldsymbol{E}_{\text{pos}} = \{1, 2, \ldots, \tilde{H} \times \tilde{W}\}$. We then send $\ddot{\boldsymbol{F}}_{S,i}$ into the multiple transformer blocks to perform the feature reconstruction task. A transformer block consists of a multi-head self-attention (MHSA) module and a multi-layer perceptron (MLP). The MHSA contains h heads each with the dimension of $d = C_D/h$. $\ddot{\boldsymbol{F}}_{S,i}$ is transformed into three groups of matrices of the query $\boldsymbol{Q}$, key $\boldsymbol{K}$, and value $\boldsymbol{V}$ through three different linear projection layers, where $\boldsymbol{Q} = [\boldsymbol{Q}_1, ..., \boldsymbol{Q}_h], \boldsymbol{K} = [\boldsymbol{K}_1, ..., \boldsymbol{K}_h], \boldsymbol{V} = [\boldsymbol{V}_1, ..., \boldsymbol{V}_h] \in \mathbb{R}^{\tilde{H}\tilde{W}\times C_D}$ for $\boldsymbol{Q}_h, \boldsymbol{K}_h, \boldsymbol{V}_h \in \mathbb{R}^{\tilde{H}\tilde{W}\times d}$. $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ are essential components of the self-attention mechanism. They are utilized to calculate the attention scores that determine how much each input element should contribute to the output. The self-attention is formulated as in:

$$\text{Atten}\left(\boldsymbol{Q}_h, \boldsymbol{K}_h, \boldsymbol{V}_h\right) = \text{Softmax}\left(\frac{\boldsymbol{Q}_h \boldsymbol{K}_h^{\text{T}}}{\sqrt{d}}\right)\boldsymbol{V}_h, \tag{12}$$

and we can obtain the output of the transformer block:

$$\begin{aligned} \text{MHSA}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) &= \text{Cat}\left(\text{Atten}\left(\boldsymbol{Q}_1, \boldsymbol{K}_1, \boldsymbol{V}_1\right), \ldots, \right. \\ &\qquad \left. \text{Atten}\left(\boldsymbol{Q}_h, \boldsymbol{K}_h, \boldsymbol{V}_h\right)\right), \\ \boldsymbol{Z}_{S,i} &= \text{MHSA}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) + \ddot{\boldsymbol{F}}_{S,i}, \\ \tilde{\boldsymbol{Z}}_{S,i} &= \text{MLP}\left(\text{Norm}\left(\boldsymbol{Z}_{S,i}\right)\right) + \boldsymbol{Z}_{S,i}, \end{aligned} \tag{13}$$

where $\text{Norm}(\cdot)$ denotes the layer normalization, $\text{Cat}(\cdot)$ indicates the concatenation operation, and $\tilde{\boldsymbol{Z}}_{S,i} \in \mathbb{R}^{\tilde{H}\tilde{W}\times C_D}$. The

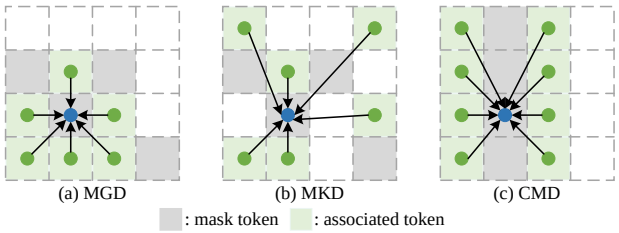


Fig. 9. Comparison of the recovery principles under different masking strategies, including MGD [20] and MKD [47] and CMD (ours).

final output of the decoder is given by:

$$\hat{\boldsymbol{Z}}_{S,i} = \mathrm{Reshape}\left(\mathrm{Linear}\left(\mathrm{TF}_N\left(\ddot{\boldsymbol{F}}_{S,i}\right)\right)\right), \tag{14}$$

where $\mathrm{TF}(\cdot)$ denotes the transformer operation explained in Eq. (13), N indicates the number of transformer blocks, and $\hat{\boldsymbol{Z}}_{S,i} \in \mathbb{R}^{C_{S,i}\times\tilde{H}\times\tilde{W}}$. $\mathrm{Linear}(\cdot)$ denotes the linear projection layer, which is utilized to restore channel numbers. $\mathrm{Reshape}(\cdot)$ transforms the feature shape from $\mathbb{R}^{\tilde{H}\tilde{W}\times C_{S,i}}$ to $\mathbb{R}^{C_{S,i}\times\tilde{H}\times\tilde{W}}$.

Finally, we restore the decoder output $\hat{\boldsymbol{Z}}_{S,i}$ to its original spatial resolution. We apply a 1×1 convolutional layer to change the channel number to $C_{S,i}\times s^2$, and then perform pixel reshuffling to adjust the spatial dimension to the same as $\boldsymbol{F}_{T,i}$. The output of CMD can be obtained as:

$$\ddot{\boldsymbol{Z}}_{S,i} = \mathrm{PR}\left(\mathrm{Conv}_{1\times1}\left(\hat{\boldsymbol{Z}}_{S,i}\right)\right), \tag{15}$$

where $\mathrm{PR}(\cdot)$ indicates the pixel reshuffle operation and $\ddot{\boldsymbol{Z}}_{S,i} \in \mathbb{R}^{C_i\times H_i\times W_i}$.

The cosine distance is adopted to calculate the distillation loss for CMD:

$$\mathcal{L}_{\mathrm{CMD}} = \sum_{i=1}^{K_S}\left(1 - \left\langle \ddot{\boldsymbol{Z}}_{S,i}, \boldsymbol{F}_{T,i}\right\rangle\right). \tag{16}$$

Note that the decoder is only used during training to perform the KD process, while only the student model is applied in the testing stage.

3) Recovery Principles under Different Masking Strategies: Applying masking strategies in the KD process is a common practice. To further highlight the effectiveness of our proposed CMD module, we present the recovery process of student feature maps under different masking strategies in Fig. 9, including CMD, MGD [20] and MKD [47].

MGD is a typical random-masking strategy, which randomly masks pixels of the student feature maps during the forward process of the backbone network. The information on the masked tokens has been 'leaked' due to the full image input. Consequently, MGD tends to extract the knowledge from adjacent tokens to restore the features of the masked regions, leading to incomplete feature recovery. On the other hand, MKD applies masks not only to the initial input image but also to the intermediate outputs throughout the model's layers. This strategy effectively prevents information leakage from masked areas, encouraging the student model to utilize a wider range of data for learning. However, the random nature of masking in both MGD and MKD disrupts the structural coherence within the time-frequency domain of spectrograms. Additionally, this randomness often causes critical data points to be overshadowed by less relevant information, which complicates the student's ability to identify and reconstruct vital features accurately.

In contrast, CMD offers two significant advantages over these random masking strategies: (1) CMD respects the inherent structure of acoustic signals, which often exhibit crucial patterns over time and across frequencies. This preservation of key characteristics enhances the feature imitation process during distillation. (2) CMD can naturally simulate interference in the marine environment. By masking specific frequency information, CMD can replicate transient noises or distortions. Furthermore, intermittently obscuring time information helps the model adapt to various lengths of silent intervals. These merits improve both the robustness and generalization capabilities of the student model (see Section IV-E).

TABLE I
CONFIGURATIONS OF DEEPSHIP AND SHIPSEAR DATASETS.

Configurations	DeepShip	ShipsEar
Dataset size	53155	3738
Category	4	5
Location	Canada Strait of Georgia	Spanish Atlantic Coast
Equipment	icListen AF Hydrophone	digitalHyd SR-1

E. Overall Loss of UDKD

To summarize, the overall loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{task}} + \lambda\mathcal{L}_{\mathrm{DBD}} + \gamma\mathcal{L}_{\mathrm{CMD}}, \tag{17}$$

where $\mathcal{L}_{\mathrm{task}}$ is the classification loss, i.e., CrossEntropy (CE) loss. λ and γ are the weights to balance the distillation losses. We perform end-to-end training and optimization. During the test phase, we discard all the attached components, and thus no extra inference costs compared with the original student model.

IV. EXPERIMENTS

A. Experiment Setup

Benchmark Datasets. We conduct comprehensive experiments on the two authentic underwater acoustic datasets, i.e., DeepShip [45] and ShipsEar [46]. Specifically, DeepShip dataset comprises four categories: Cargo, Passenger-boats, Tug, and Tanker. ShipsEar dataset groups underwater acoustic records into five categories: Class-A (Fishing boats, Trawlers, Mussel boats, Tugboats, and Dredgers), Class-B (Motorboats, Pilot boats, and Sailboats), Class-C (Passenger ferries), Class-D (Ocean liners and RO-RO vessels), and Class-E (Background noise). Following the standard practice [1], we convert all the audio recordings in both datasets into Mel spectrograms, each with 3 seconds recording with 50% overlapping. Details of the two datasets are provided in Table I.

Implementation Details. We train the UDKD for up to 30 epochs using the AdamW optimizer with a batch size of 16. The initial learning rate is 0.001 and it is adjusted

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT KD METHODS ON DEEPSHIP AND SHIPSEAR DATASETS. TOP THREE RESULTS ARE COLORED IN RED, BLUE, AND GREEN, RESPECTIVELY. ↑ INDICATES THE PERFORMANCE IMPROVEMENT OF UDKD COMPARED WITH THE SECOND-BEST METHOD. ⇑ DENOTES THE PERFORMANCE IMPROVEMENT OF UDKD COMPARED WITH THE STUDENT MODEL.

Distillation Mechanism	Method	DeepShip ACC	DeepShip F1-Score	DeepShip AUC	ShipsEar ACC	ShipsEar F1-Score	ShipsEar AUC
Baseline (No KD)	Teacher (ResNet-18)	83.88	83.63	96.16	97.74	97.73	99.91
	Student (CNN-3)	73.50	73.48	90.59	84.31	84.22	97.46
Logit	KD [9]	79.88	79.95	94.15	90.29	90.24	98.82
	CC [33]	76.13	76.14	92.00	86.70	86.63	98.22
	DKD [36]	79.63	79.70	94.00	90.91	90.91	99.21
	WSLD [21]	79.00	79.06	94.62	91.62	91.58	99.14
	SRRL [48]	78.38	78.43	93.56	90.51	90.51	98.78
	NKD [49]	74.63	75.00	92.31	90.03	89.97	99.03
	LSKD [50]	76.88	76.90	93.22	90.96	90.91	99.15
	SDD [18]	77.00	77.03	93.76	91.22	91.18	98.96
Feature	AT [17]	79.25	79.31	94.19	87.85	87.83	98.60
	SP [38]	79.75	79.82	94.23	87.63	87.57	98.54
	RKD [14]	79.50	79.57	93.85	90.51	90.51	99.00
	PKT [51]	79.38	79.44	94.11	88.83	88.77	98.31
	FSP [52]	78.63	78.68	94.09	90.69	90.64	99.21
	NST [53]	79.25	79.31	94.02	90.78	90.78	98.55
	VID [54]	79.38	79.44	94.05	89.36	89.30	99.02
	ICKD [55]	77.38	77.41	93.93	86.66	86.63	97.81
	MGD [20]	79.38	79.44	94.05	91.09	91.04	98.70
	MKD [47]	77.88	77.92	93.67	90.03	89.97	98.58
	CAT-KD [56]	78.13	78.17	93.37	90.56	90.51	98.87
	UATR-KD [57]	77.63	77.66	93.88	87.90	87.83	98.35
	UDKD (Ours)	81.00	81.09	94.91	94.81	94.79	99.50
	↑	+ 1.12	+ 1.14	+ 0.29	+ 3.19	+ 3.21	+ 0.29
	⇑	+ 7.50	+ 7.61	+ 4.32	+ 10.5	+ 10.6	+ 2.04

by the Cosine Annealing Warm Restarts strategy. The hyper-parameter r decreases gradually with the reduction of the spatial dimension in the feature map, i.e., $r = \{4, 2, 1\}$. Our decoder consists of 4 transformer blocks, where each transformer block is configured with 256 channels and 8 heads. The masking ratio is set to 0.1 by default. We choose $\lambda = 1$ for DBD loss and $\gamma = 1$ for CMD loss. For each dataset, 70% samples are randomly selected for training, 10% samples are randomly selected for validation, and the remaining 20% samples are utilized for testing. We repeat the experiments 10 times with different splits of the training, validation and test sets, and the average of the 10 experimental results are reported. We utilize the PyTorch 1.13.0 framework to build the entire model architecture and perform all experiments on an NVIDIA GeForce RTX 3080.

Evaluation Metrics. We evaluate the performance by Accuracy (ACC), F1-Score, and Area Under Curve (AUC). ACC measures the proportion of correctly classified instances out of the total instances. F1-Score provides a balance between precision and recall, especially when the dataset is imbalanced. AUC comprehensively measures the effectiveness across all the possible classification thresholds. It represents the probability that the model ranks a randomly chosen positive sample higher than a randomly chosen negative sample.

B. Comparison With the State-of-the-arts

Our proposed UDKD is compared with 20 classical or popular state-of-the-art (SOTA) knowledge distillation methods, including KD [9], CC [33], DKD [36], WSLD [21], SRRL [48], NKD [49], LSKD [50], SDD [18], AT [17], SP [38], RKD [14], PKT [51], FSP [52], NST [53], VID [54], ICKD [55], MGD [20], MKD [47], CAT-KD [56], and UATR-KD [57]. Among them, the first 8 methods (i.e., KD, CC, DKD, WSLD, SRRL, NKD, LSKD, SDD) belong to the logit-based distillation. The other 12 methods (i.e., AT, SP, RKD, PKT, FSP, NST, VID, ICKD, MGD, MKD, CAT-KD, UATR-KD) belong to the feature-based distillation. From Table II, it is observed that UDKD significantly improves the performance of student and outperforms the SOTA methods substantially. Specifically, on the ShipsEar dataset, UDKD improves the student with 10.5% ACC, which greatly fills the performance gap between the student and the teacher. Besides, compared with the second-best method, UDKD brings 3.19% ACC improvement. These observations strongly demonstrate the effectiveness and superiority of learning low/high-frequency features and time-frequency distribution patterns in the UATR tasks. By comparison, many logit-based and feature-based distillations only achieve a moderate performance. For example, MGD [20] masks random pixels in the student's feature and forces it to generate the teacher's entire feature representation through a convolutional block. This disrupts the structural integrity of information in the time-frequency domain. In contrast, our masking strategy in CMD is more suitable for analyzing underwater acoustic knowledge.

C. Ablation Study and Hyper-parameter Analysis

We conduct extensive ablation studies and hyper-parameter analysis to examine the effectiveness of UDKD. All the experiments are performed on DeepShip and ShipsEar datasets.

TABLE III
OVERALL ABLATION STUDY OF UDKD.

KD Types	DeepShip ACC	ShipsEar ACC
-	73.50	84.31
DBD	75.88	90.69
CMD	76.63	89.76
Low, CMD	79.63	93.75
High, CMD	78.25	93.35
DBD, CMD	**81.00**	**94.81**

TABLE IV
PERFORMANCE EVALUATION OF THE HYPER-PARAMETER r.

r	DeepShip ACC	ShipsEar ACC
1	79.25	93.62
2	78.13	92.95
4	78.75	92.55
$\{4, 2, 1\}$	**81.00**	**94.81**

TABLE V
PERFORMANCE EVALUATION OF THE FILTER SHAPE.

Filter Shape	DeepShip ACC	ShipsEar ACC
Circle	**81.00**	**94.81**
Square	79.25	93.09
Rhombus	78.63	93.22
Irregular	76.13	88.65

TABLE VI
PERFORMANCE EVALUATION OF THE MASKING RATIO.

Mask Ratio	DeepShip ACC	ShipsEar ACC
0	78.63	91.49
0.1	**81.00**	**94.81**
0.3	80.13	93.75
0.5	79.88	93.22
0.7	78.38	92.15
0.9	75.88	90.16

TABLE VII
PERFORMANCE EVALUATION OF THE MASKING STRATEGY.

Mask Strategy	DeepShip ACC	ShipsEar ACC
Freq	79.63	93.75
Time	78.13	93.62
Random	78.50	92.29
Block	76.63	89.72
Grid	75.88	88.83
Full-freq/time	**81.00**	**94.81**

TABLE VIII
PERFORMANCE EVALUATION OF THE DECODER DEPTH.

Decoder Depth	DeepShip ACC	ShipsEar ACC
2	80.50	93.62
4	**81.00**	**94.81**
6	80.63	94.02
8	80.25	94.41

TABLE IX
PERFORMANCE EVALUATION OF THE DECODER WIDTH.

Decoder Width	DeepShip ACC	ShipsEar ACC
128	79.50	93.88
256	**81.00**	**94.81**
512	80.13	94.28

TABLE X
PERFORMANCE EVALUATION OF THE DECODER HEAD.

Decoder Head	DeepShip ACC	ShipsEar ACC
4	79.38	93.09
8	**81.00**	**94.81**
16	80.38	93.75

TABLE XI
PERFORMANCE EVALUATION OF THE LOSS WEIGHT λ.

λ	DeepShip ACC	ShipsEar ACC
1	**81.00**	**94.81**
3	79.88	94.15
5	79.13	94.41
7	80.63	94.55

TABLE XII
PERFORMANCE EVALUATION OF THE LOSS WEIGHT γ.

γ	DeepShip ACC	ShipsEar ACC
1	**81.00**	**94.81**
3	80.63	93.88
5	79.75	94.02
7	79.88	94.28

TABLE XIII
PERFORMANCE EVALUATION OF THE LOSS FUNCTION.

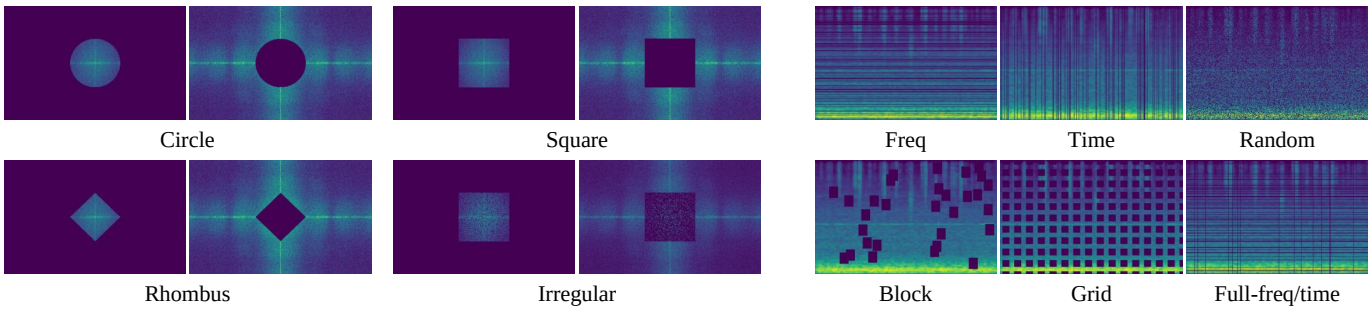
Loss Function	DeepShip ACC	ShipsEar ACC
MSE	78.50	90.43
Cosine Similarity	**81.00**	**94.81**

Unless specified, the experimental settings remain the same as those explained in "Implementation Details" in Sec. IV-A.

Overall ablations. The proposed UDKD consists of two essential components, i.e., Dual-frequency Band Distillation (DBD) and Cross-domain Masked Distillation (CMD). We first examine the individual contribution of each component. The ablation results are presented in Table III, where Low and High indicate that we only use low-frequency components and high-frequency components in DBD, respectively. In Table III, we can find out that without any KD process, the student achieves low prediction accuracy. When adding DBD or CMD, the student achieves 6.38% and 5.45% improvements on ShipsEar dataset, respectively. By further analysis, we can find out that adding Low and High on the basis of CMD enables the student to achieve the performance gain, i.e., 3.99% and 3.59% on ShipsEar dataset, respectively. Finally, combining DBD and CMD can further improve the performance gain, which strongly demonstrates that all the designed KD components are effective and contribute to the overall performance.

Hyper-parameter r. Table IV studies the effects of hyper-parameter r in DBD, which determines the number of frequencies processed by the low/high-pass filters. A larger r includes more frequencies in the low-pass and fewer in the high-pass filter. It may be straightforward to set r as a constant. However, as the model goes deeper, a static value of r will lead to disproportionate filtering effects across layers. For instance, with $r = 4$, a 56×56 feature map processes 45 frequencies in the low-pass and 3091 in high-pass filter. For a 28×28 feature map, the low-pass still processes 45 frequencies, but the high-pass only handles 739 frequencies. Obviously, this disrupts the balance between the low and high frequencies, restraining the efficiency of the KD process. Thus, we propose dynamically adjusting r in proportion to changes in the spatial scale of feature maps, i.e., $r = \{4, 2, 1\}$. As a result, both low and high-pass filters can adapt their frequency processing capabilities accordingly. The results in Table IV confirm our proposal.

Filter shape. We compare different filter shapes in DBD



(a) Visualization of filter shapes (b) Visualization of masking strategies

Fig. 10. (a) Visualization of filter shapes. (b) Visualization of masking strategies. In order to improve observability, we increase the size of the filter shape areas and adjusted the masking ratio in the visualizations. In (a), for each filter shape, the left column denotes the low-pass filter and right column indicates the high-pass filter.

TABLE XIV
TEACHER-STUDENT OUTPUT LOGIT DIFFERENCES ON DEEPSHIP AND SHIPSEAR DATASETS.

	CC [33]	DKD [36]	WSLD [21]	SRRL [48]	NKD [49]	LSKD [50]	SP [38]	RKD [14]	PKT [51]	FSP [52]	VID [54]	ICKD [55]	MGD [20]	UATR-KD [57]	**UDKD (Ours)**
DeepShip	7.05	5.80	5.47	5.53	6.34	5.47	5.73	5.77	5.75	5.81	5.83	5.84	5.60	5.22	**5.06**
ShipsEar	3.39	2.16	2.27	2.76	2.93	2.25	3.27	2.20	2.69	2.32	2.51	4.06	2.74	3.05	**1.67**

and the results are shown in Table V. Fig. 10(a) shows the different filter shapes that focus on different directions [22]. Specifically, square shapes focus on horizontal and vertical directions, rhombus shapes emphasize more frequencies on the diagonal, and irregular shapes randomly affect some directions. The results demonstrate that the circle shapes with equal attention to all directions of frequencies perform favorably.

Masking ratio. We investigate the influence of different masking ratios. As show in Table VI, with the increase of masking ratio, ACC gradually decreases, which fits our intuitive impression. The larger masking ratio makes the distillation process more difficult. However, it is worth noting that even with a masking ratio as high as 90%, we still achieve a 5.85% improvement on the ShipsEar dataset compared with student without KD. This verifies the effectiveness of our distillation mechanism. Additionally, when the masking ratio is set to 0, CMD degenerates into the basic autoencoder architecture. Due to the interference of redundant information in the Mel spectrograms, the teacher transfers ineffective knowledge to the student, resulting in performance degradation.

Masking strategy. We further examine how the masking strategies influence the distillation performance. Fig. 10(b) shows the different masking strategies, and we conduct these strategies in the CMD module, including random masking, block masking, and grid masking. The experimental results are presented in Table VII, where Freq and Time mean that our designed masking strategy is only performed on specific time slots and frequency points. As observed, the designed masking strategy earns the best performance, which strongly demonstrates its effectiveness. In comparison, the block and grid masking disrupt the structural integrity and continuity of the spectral knowledge in the time-frequency domain, rendering to a moderate performance.

Decoder design. A well-designed decoder is important for reconstruction tasks [47]. Therefore, we explore different options for decoder design, as studied in Table VIII, Table IX, and Table X. Specifically, we change the depth, width (number of channels), and the number of heads of the decoder. When we change one hyperparameter, the others remain at their default values described in "Implementation Details" in Sec. IV-A. The experimental results show that a decoder with a depth of 4, a width of 256, and 8 heads achieves superior performance.

Loss weights. In Table XI and Table XII, we investigate the effects of loss weights λ and γ on KD performance. For these two loss weights, when we change the value of one, the value of the other is set to 1. As observed, the performance gap between the worst and best is within 0.66% ACC. This result demonstrates that the student performance is not sensitive to these loss weights.

Loss function. In UDKD, we utilize cosine similarity to calculate the feature distillation losses. We herein explore the effects of other loss functions on KD performance. We replace the cosine similarity with MSE for the experiments and the results are shown in Table XIII. As observed, cosine similarity achieves better performance.

D. Teacher-Student Differences Analysis

Better learning in-distribution knowledge of teacher can help reduce the performance gap between the student and the teacher. Therefore, we investigate whether UDKD-trained student indeed better captures the in-distribution knowledge of teacher. The in-distribution knowledge can be effectively represented by output logit [58]. Thus we calculate the teacher-student output logit differences by mean-square error. As shown in Table XIV, the output logit differences of UDKD are consistently smaller than other KD methods, which indicates that the in-distribution knowledge of student in UDKD is closer to the teacher. It is worth noting that UDKD does not

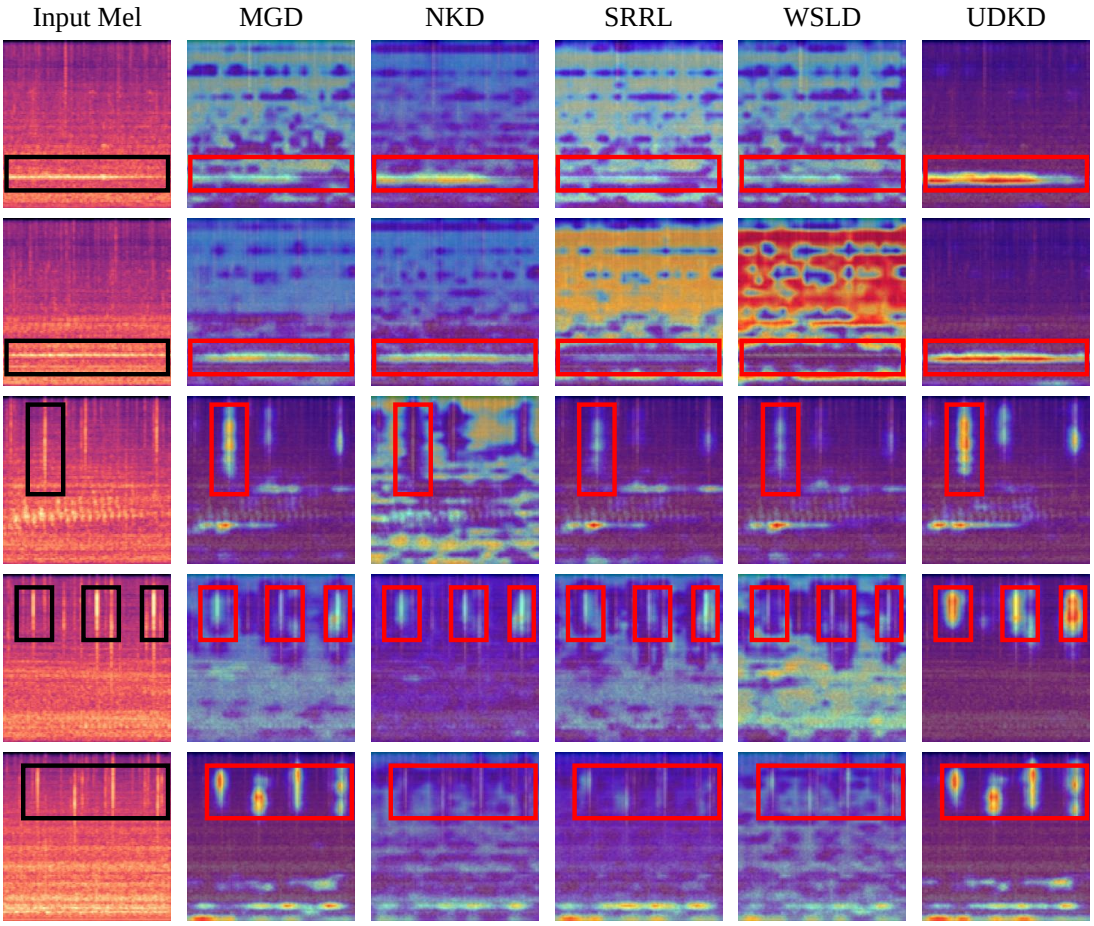


Fig. 11. Visualization of attention maps of student with different KD paradigms on ShipsEar dataset.

TABLE XV
ACC UNDER FGSM WHITE-BOX ATTACK WITH VARIOUS PERTURBATION WEIGHTS ε ON SHIPSEAR DATASET. ε CONTROLS THE DIFFERENCE BETWEEN ORIGINAL SAMPLE AND ADVERSARIAL SAMPLE, WHERE A LARGER ε INDICATES A MORE SIGNIFICANT DIFFERENCE.

Method	$\varepsilon = 0.0001$	$\varepsilon = 0.0005$	$\varepsilon = 0.001$
KD [9]	86.30	68.84	41.13
DKD [36]	88.79	71.85	40.60
WSLD [21]	88.70	70.92	41.89
NKD [49]	85.59	70.39	39.94
LSKD [50]	89.10	71.45	43.79
SDD [18]	88.30	71.32	43.26
AT [17]	85.06	69.33	45.66
SP [38]	84.57	69.06	42.77
RKD [14]	86.66	69.73	32.93
PKT [51]	87.50	70.04	42.02
MGD [20]	87.50	73.23	42.91
CAT-KD [56]	86.84	68.97	38.12
UDKD (Ours)	**92.42**	**75.00**	**45.97**

explicitly compute logit-based losses, while our output logit differences are still minimal.

E. Robustness Evaluation

In real underwater environments, acoustic signals are inevitably disturbed by underwater environmental noise and multi-path propagation. Therefore, it is important to evaluate the robustness of our distillation framework. Adversarial learning can easily deceive a model by adding small but deliberate worst-case perturbations that are difficult to detect in the input image. Thus we conduct adversarial experiments to evaluate the model robustness. Specifically, we compare the robustness of KD mechanisms under FGSM white-box attack [59]. As shown in Table XV, UDKD significantly improves robustness and consistently outperforms other KD methods under different perturbations. The results show that UDKD has better potential in complex ocean environments.

TABLE XVI
PERFORMANCE EVALUATION OF THE LOW/HIGH-FREQUENCY FEATURES.

Data Type	DeepShip (ACC)	ShipsEar (ACC)
Low-pass spectra	78.0	92.3
High-pass spectra	54.9	82.1
Mel spectrograms	**83.9**	**97.7**

F. Discussion

We present a detailed discussion from two perspectives to further analyze our proposed UDKD.

Low/high-frequency features. In DBD module, the student absorbs the low/high-frequency knowledge of the teacher. We are therefore interested in their individual effectiveness. To

TABLE XVII
PERFORMANCE EVALUATION OF THE TRANSFORMER-BASED DECODER AND CNN-BASED DECODER.

Decoder	DeepShip (ACC)	ShipsEar (ACC)
CNN-based decoder	76.38	86.84
Transformer-based decoder	**81.00**	**94.81**

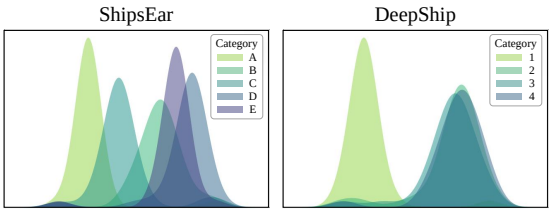


Fig. 12. Distribution of the high-frequency information for the DeepShip and ShipsEar datasets, where categories 1, 2, 3, and 4 in the DeepShip dataset represent the Cargo, Passenger boats, Tug, and Tanker classes, respectively.

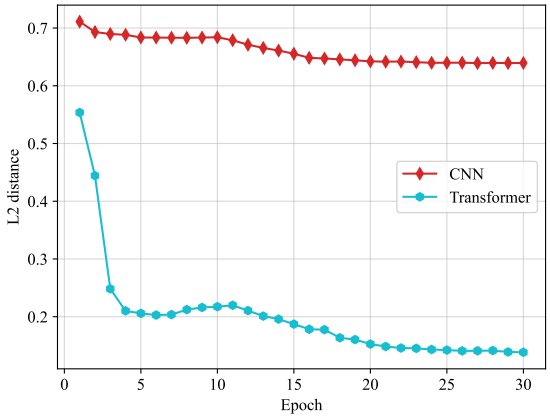


Fig. 13. The L2 distance between the reconstructed student's feature maps and the corresponding teacher's feature maps on ShipsEar dataset.

this end, we perform the low/high-pass filtering on the Mel spectrograms on the DeepShip and ShipsEar datasets, and generate the corresponding low/high-pass spectra. We utilize the ResNet-18 to train these spectra, and the results are shown in Table XVI. For the ShipsEar dataset, the low/high-pass spectra achieve close performance to the Mel spectrograms. However, the accuracy of the high-pass spectra decreases significantly on the DeepShip dataset. This can be attributed to the low separability of the high-frequency features in the DeepShip dataset. As shown in Fig. 12, in DeepShip, the distribution of the high-frequency information for the last three categories is highly overlapping, while the ShipsEar dataset has a large gap in the distribution of the high-frequency information for each category. In summary, the low/high-frequency features work in synergy to enhance the student's learning.

Transformer-based decoder v.s. CNN-based decoder. A well-performing decoder is important for the feature reconstruction. The decoder in MGD [20] consists of multiple convolutional layers, while our proposed CMD adopts the transformer-based decoder. Therefore, we compare the performance between different types of decoders through two schemes. We first calculate the L2 distance between the reconstructed student's feature maps and the complete teacher's feature maps during the training process, and the results on ShipsEar dataset are shown in Fig. 13. Throughout the training process, the L2 distance of the transformer-based decoder is significantly smaller than that of the CNN-based decoder. We further replace the transformer-based decoder in CMD with the CNN-based decoder to verify the advantages of transformer-based decoder, where the CNN-based decoder consists of three BasicBlocks (the building blocks in the ResNet architecture). As observed in Table XVII, the transformer-based decoder achieves superior performance, and thus it is more conducive to feature reconstruction.

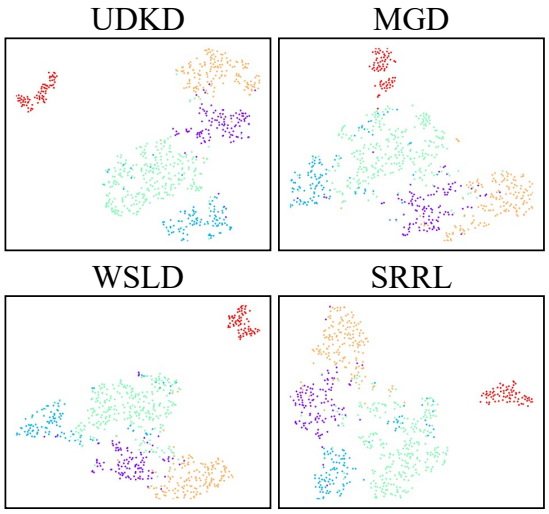


Fig. 14. t-SNE of features learned by our proposed UDKD, MGD [20], WSLD [21], and SRRL [48] on ShipsEar dataset.

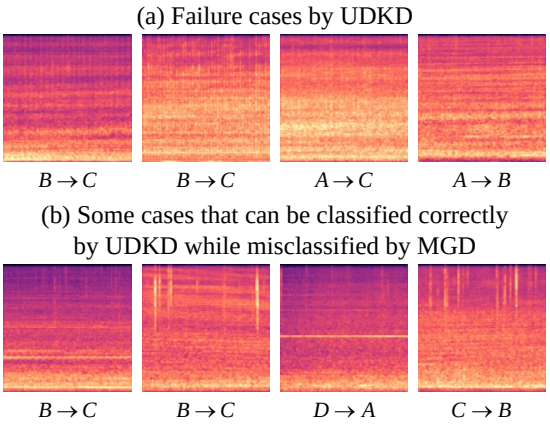


Fig. 15. (a) Some failure cases from the proposed UDKD. $a \rightarrow b$ indicates that true label is a and prediction is b. (b) Example cases that can be classified correctly by the UDKD-trained student while misclassified by the MGD-trained student [20]. $a \rightarrow b$ denotes that UDKD predicts a, but MGD predict b. We report the results on ShipsEar dataset.

G. Qualitative Analysis

We present visualizations from three perspectives to illustrate the principles of our approach.

Visualization of attention maps. Fig. 11 shows the attention maps of student with different KD frameworks. According

to the feature comparison in the red box, UDKD shows the high responses in some line spectra and time periods, while the attention distribution of other methods is more chaotic and the responses are indistinctive. This demonstrates that the student using UDKD learns more representative knowledge from the teacher than other feature-based and logit-based methods.

t-SNE visualizations. We visualize the inter-class distance map via t-SNE projection in Fig. 14. We can observe that the representations of UDKD are more separable than other KD mechanisms, showing the proposed UDKD can enhance the discriminative capacity of student.

Failure cases. Although the proposed UDKD can achieve optimal performance in the UATR tasks, there are still some challenging scenarios. As shown in Fig. 15(a), it is challenging for UDKD to classify Mel spectrograms without obvious line spectra and time periods. The time-frequency distribution patterns of these Mel spectrograms change slowly, making it difficult for UDKD to extract effective features from them. Furthermore, we visualize some cases that can be classified correctly by the UDKD-trained student while misclassified by the student trained with MGD [20], and the results are shown in Fig. 15(b). From this figure, we can observe that the samples misclassified by MGD all contain distinct line spectra and time regions. These results verify our proposal that UDKD can help the student extract the low/high-frequency features, thereby improving its discrimination ability to the acoustic signals.

V. Conclusion

This paper reveals the challenges of the feature-based knowledge distillation paradigms in the underwater acoustic target recognition (UATR). To overcome these challenges, we propose the Union-Domain Knowledge Distillation (UDKD), which employs two new strategies of knowledge transferring for acoustic data. Specifically, we introduce the Dual-frequency Band Distillation (DBD) module that enables the student to efficiently learn the low/high-frequency features in the frequency domain spectra. Furthermore, we design the Cross-domain Masked Distillation (CMD) module that guides the student to learn the intricate time-frequency distribution patterns of acoustic signals. Extensive experiments on two benchmark datasets demonstrate the effectiveness of UDKD. The proposed union-domain solution offers a new perspective for the community to rethink the model design of UATR tasks.

References

[1] A. Zhou, X. Li, W. Zhang, D. Li, K. Deng, K. Ren, and J. Song, "A Novel Cross-Attention Fusion-Based Joint Training Framework for Robust Underwater Acoustic Signal Recognition," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[2] S. Tian, D. Chen, H. Wang, and J. Liu, "Deep Convolution Stack for Waveform in Underwater Acoustic Target Recognition," *Scientific reports*, vol. 11, no. 1, p. 9614, 2021.

[3] F. Liu, T. Shen, Z. Luo, D. Zhao, and S. Guo, "Underwater target recognition using convolutional recurrent neural networks with 3-D Mel-spectrogram and data augmentation," *Applied Acoustics*, vol. 178, p. 107989, 2021.

[4] Y. Xie, J. Ren, and J. Xu, "Underwater-art: Expanding information perspectives with text templates for underwater acoustic target recognition," *The Journal of the Acoustical Society of America*, vol. 152, no. 5, pp. 2641–2651, 2022.

[5] X. Cao, L. Ren, and C. Sun, "Research on Obstacle Detection and Avoidance of Autonomous Underwater Vehicle Based on Forward-Looking Sonar," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[6] P. Zhu, Y. Zhang, Y. Huang, B. Lin, M. Zhu, K. Zhao, and F. Zhou, "SFC-Sup: Robust Two-Stage Underwater Acoustic Target Recognition Method Based On Supervised Contrastive Learning," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[7] Y. Wang, C. Tang, S. Wang, L. Cheng, R. Wang, M. Tan, and Z. Hou, "Target Tracking Control of a Biomimetic Underwater Vehicle Through Deep Reinforcement Learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 8, pp. 3741–3752, 2021.

[8] Y. Shang, Z. Yuan, B. Xie, B. Wu, and Y. Yan, "Post-Training Quantization on Diffusion Models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1972–1981.

[9] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," *Computer Science*, vol. 14, no. 7, pp. 38–39, 2015.

[10] Z. Pan, J. Cai, and B. Zhuang, "Stitchable Neural Networks," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16 102–16 112, 2023.

[11] G. Fang, X. Ma, M. Song, M. B. Mi, and X. Wang, "DepGraph: Towards Any Structural Pruning," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16 091–16 101, 2023.

[12] B. Zhao, R. Song, and J. Liang, "Cumulative Spatial Knowledge Distillation for Vision Transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 6146–6155.

[13] C. Yang, Z. An, H. Zhou, L. Cai, X. Zhi, J. Wu, Y. Xu, and Q. Zhang, "MixSKD: Self-Knowledge Distillation from Mixup for Image Recognition," in *European Conference on Computer Vision*. Springer, 2022, pp. 534–551.

[14] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational Knowledge Distillation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3967–3976.

[15] D. Chen, J.-P. Mei, Y. Zhang, C. Wang, Z. Wang, Y. Feng, and C. Chen, "Cross-Layer Distillation with Semantic Calibration," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 8, 2021, pp. 7028–7036.

[16] S. Shin, J. Lee, J. Lee, Y. Yu, and K. Lee, "Teaching Where to Look: Attention Similarity Knowledge Distillation for Low Resolution Face Recognition," in *European Conference on Computer Vision*. Springer, 2022, pp. 631–647.

[17] S. Zagoruyko and N. Komodakis, "Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer," in *International Conference on Learning Representations*, 2017.

[18] S. Wei, C. Luo, and Y. Luo, "Scale Decoupled Distillation," *ArXiv*, vol. abs/2403.13512, 2024.

[19] M. Kang, J. Zhang, J. Zhang, X. Wang, Y. Chen, Z. Ma, and X. Huang, "Alleviating Catastrophic Forgetting of Incremental Object Detection via Within-Class and Between-Class Knowledge Distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18 894–18 904.

[20] Z. Yang, Z. Li, M. Shao, D. Shi, Z. Yuan, and C. Yuan, "Masked Generative Distillation," in *European Conference on Computer Vision*. Springer, 2022, pp. 53–69.

[21] H. Zhou, L. Song, J. Chen, Y. Zhou, G. Wang, J. Yuan, and Q. Zhang, "Rethinking Soft Labels for Knowledge Distillation: A Bias Variance Tradeoff Perspective," in *International Conference on Learning Representations*, 2021.

[22] J. Xie, W. Li, X. Zhan, Z. Liu, Y.-S. Ong, and C. C. Loy, "Masked Frequency Modeling for Self-Supervised Visual Pre-Training," in *The Eleventh International Conference on Learning Representations*, 2023.

[23] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "SimMIM: A Simple Framework for Masked Image Modeling," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 9653–9663.

[24] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer, "Masked Feature Prediction for Self-Supervised Visual Pre-Training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 668–14 678.

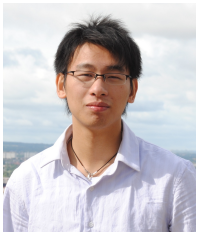
[25] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT Pre-Training of Image Transformers," in *International Conference on Learning Representations*, 2022.

[26] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[27] O. Adam, "The use of the Hilbert-Huang transform to analyze transient signals emitted by sperm whales," *Applied Acoustics*, vol. 67, no. 11, pp. 1134–1143, 2006.
[28] J. Seok and K. Bae, "Target Classification Using Features Based on Fractional Fourier Transform," *IEICE Transactions on Information and Systems*, vol. E97-D, no. 9, p. 2518 – 2521, 2014.
[29] T. Lim, K. Bae, C. Hwang, and H. Lee, "Classification of underwater transient signals using MFCC feature vector," in *2007 9th International Symposium on Signal Processing and Its Applications*, 2007, pp. 1–4.
[30] Y. Sun, Y. Peng, L. Mu, F. Zhang, and L. Cao, "Classification Technology of Underwater Targets in Decision Tree Based on Different Measurement Criteria," in *2021 IEEE International Conference on Signal Processing, Communications and Computing*, 2021, pp. 1–4.
[31] H. Li, Y. Cheng, W. Dai, and Z. Li, "A method based on wavelet packets-fractal and SVM for underwater acoustic signals recognition," in *2014 12th International Conference on Signal Processing*, 2014, pp. 2169–2173.
[32] T. Gao, Q. Niu, J. Zhang, T. Chen, S. Mei, and A. Jubair, "Global to Local: A Scale-Aware Network for Remote Sensing Object Detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.
[33] B. Peng, X. Jin, J. Liu, D. Li, Y. Wu, Y. Liu, S. Zhou, and Z. Zhang, "Correlation Congruence for Knowledge Distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5007–5016.
[34] G. Xu, Z. Liu, X. Li, and C. C. Loy, "Knowledge Distillation Meets Self-supervision," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 588–604.
[35] M. Xue, J. Song, X. Wang, Y. Chen, X. Wang, and M. Song, "KD-Explainer: A Task-oriented Attention Model for Explaining Knowledge Distillation," *ArXiv*, vol. abs/2105.04181, 2021.
[36] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang, "Decoupled Knowledge Distillation," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2022, pp. 11 953–11 962.
[37] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for Thin Deep Nets," *CoRR*, vol. abs/1412.6550, 2014.
[38] F. Tung and G. Mori, "Similarity-Preserving Knowledge Distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1365–1374.
[39] P. Chen, S. Liu, H. Zhao, and J. Jia, "Distilling Knowledge via Knowledge Review," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2021, pp. 5008–5017.
[40] B. Heo, J. Kim, S. Yun, H. Park, N. Kwak, and J. Y. Choi, "A Comprehensive Overhaul of Feature Distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2019.
[41] C. Schörkhuber and A. Klapuri, "Constant-Q transform toolbox for music processing," *Proc. 7th Sound and Music Computing Conf.*, 01 2010.
[42] B. McFee, C. Raffel, D. Liang, D. Ellis, M. Mcvicar, E. Battenberg, and O. Nieto, "librosa: Audio and Music Signal Analysis in Python," 01 2015, pp. 18–24.
[43] J. Hong, M. Kim, J. Y. Choi, and Y. M. Ro, "Watch or Listen: Robust Audio-Visual Speech Recognition with Visual Corruption Modeling and Reliability Scoring," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18 783–18 794, 2023.
[44] X. Zhang, S. Li, X. Li, P.-C. Huang, J. Shan, and T. Chen, "Destseg: Segmentation guided denoising student-teacher for anomaly detection," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3914–3923, 2022.
[45] M. Irfan, Z. Jiangbin, S. Ali, M. Iqbal, Z. Masood, and U. Hamid, "DeepShip: An Underwater Acoustic Benchmark Dataset and a Separable Convolution Based Autoencoder for Classification," *Expert Systems with Applications*, vol. 183, p. 115270, 2021.
[46] D. Santos-Domínguez, S. Torres-Guijarro, A. Cardenal-López, and A. Pena-Gimenez, "ShipsEar: An underwater vessel noise database," *Applied Acoustics*, vol. 113, pp. 64–69, 2016.
[47] S. Lao, G. Song, B. Liu, Y. Liu, and Y. Yang, "Masked Autoencoders Are Stronger Knowledge Distillers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 6384–6393.
[48] J. Yang, B. Martinez, A. Bulat, and G. Tzimiropoulos, "Knowledge Distillation Via Softmax Regression Representation Learning," in *International Conference on Learning Representations*, 2020.
[49] Z. Yang, A. Zeng, Z. Li, T. Zhang, C. Yuan, and Y. Li, "From Knowledge Distillation to Self-Knowledge Distillation: A Unified Approach with Normalized Loss and Customized Soft Labels," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 185–17 194.
[50] S. Sun, W. Ren, J. Li, R. Wang, and X. Cao, "Logit Standardization in Knowledge Distillation," *ArXiv*, vol. abs/2403.01427, 2024.
[51] N. Passalis, M. Tzelepi, and A. Tefas, "Probabilistic Knowledge Transfer for Deep Representation Learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 5, pp. 2030–2039, 2020.
[52] J. Yim, D. Joo, J. Bae, and J. Kim, "A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4133–4141.
[53] Z. Huang and N. Wang, "Like What You Like: Knowledge Distill via Neuron Selectivity Transfer," *CoRR*, vol. abs/1707.01219, 2017.
[54] S. Ahn, S. X. Hu, A. Damianou, N. D. Lawrence, and Z. Dai, "Variational Information Distillation for Knowledge Transfer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9163–9171.
[55] L. Liu, Q. Huang, S. Lin, H. Xie, B. Wang, X. Chang, and X. Liang, "Exploring Inter-Channel Correlation for Diversity-preserved Knowledge Distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8271–8280.
[56] Z. Guo, H. Yan, H. Li, and X.-L. Lin, "Class Attention Transfer Based Knowledge Distillation," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11 868–11 877, 2023.
[57] S. Yang, A. Jin, X. Zeng, H. Wang, X. Hong, and M. Lei, "Underwater acoustic target recognition based on knowledge distillation under working conditions mismatching," *Multimedia Systems*, vol. 30, no. 1, pp. 1–14, 2024.
[58] X. Deng, J. Zheng, and Z. Zhang, "Personalized Education: Blind Knowledge Distillation," in *European Conference on Computer Vision*. Springer, 2022, pp. 269–285.
[59] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," *Computer Science*, 2014.

Xiaohui Chu (Student Member, IEEE) received the B.S. degree in software engineering from the Shanxi University, Shanxi, China, in 2019, and the M.S. degree in software engineering from the Taiyuan University of Technology, Shanxi, China, in 2022. He is currently studying for Ph.D. degree at the Beijing Institute of Technology. His current research interests include image quality assessment, underwater acoustic target recognition, and computer vision.

Haoran Duan (Member, IEEE) received a Distinction M.S. degree in Data Science from Newcastle University, UK. He obtained PhD degree in the Department of Computer Science, Durham University. He has been a research associate in Newcastle University, UK, working on deep learning applications. His current research interests focus on the applications/theories of deep learning.



Zhenyu Wen (Senior Member, IEEE) is currently a Tenure-Tracked Professor with the Institute of Cyberspace Security and the College of Information Engineering, Zhejiang University of Technology. His current research interests include the IoT, crowd sources, AI systems, and cloud computing. For his contributions to the area of scalable data management for the Internet of Things, he was awarded the IEEE TCSC Award for Excellence in Scalable Computing (Early Career Researchers) in 2020.

Lijun Xu received the M.S. and Ph.D. degrees in University of Chinese Academy of Science, China, in 2004 and 2018, respectively. He is currently a director of the ocean information technology institute in school of information and electronics, Beijing Institute of Technology. His primary research area is underwater acoustic signal processing.

Runze Hu (Member, IEEE) received the B.S. degree in computer science from North China Electric Power University, Baoding, China, in 2014, and the M.Sc. degree in computer science and the Ph.D. degree in electrical and electronics engineering from the University of Manchester, Manchester, U.K., in 2016 and 2020, respectively. His current research interests include image quality assessment, uncertainty quantification techniques, and underwater acoustic ranging.

Wei Xiang (Senior Member, IEEE) received the B.Eng. and M.Eng. degrees in electronic engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 1997 and 2000, respectively, and the Ph.D. degree in telecommunications engineering from the University of South Australia, Adelaide, Australia, in 2004. From 2004 to 2015, he was with the School of Mechanical and Electrical Engineering, University of Southern Queensland, Toowoomba, Australia. He is currently the Founding Professor and the Head of the Discipline of Internet of Things Engineering with the College of Science and Engineering, James Cook University, Cairns, Australia. He has authored or co-authored over 200 peer-reviewed journal and conference papers. His research interests are in the broad areas of communications and information theory, particularly the Internet of Things, and coding and signal processing for multimedia communications systems. He is an Elected Fellow of the IET and Engineers Australia. He received the TNQ Innovation Award in 2016, and was a finalist for 2016 Pearcey Queensland Award. He was a co-recipient of three best paper awards at 2015 WCSP, 2011 IEEE WCNC, and 2009 ICWMC. He has been awarded several prestigious fellowship titles. He was named a Queensland International Fellow (2010–2011) by the Queensland Government of Australia, an Endeavour Research Fellow (2012–2013) by the Commonwealth Government of Australia, a Smart Futures Fellow (2012–2015) by the Queensland Govern ment of Australia, and a JSPS Invitational Fellow jointly by the Australian Academy of Science and Japanese Society for Promotion of Science (2014–2015). He is the Vice Chair of the IEEE Northern Australia Section. He was an Editor of the IEEE COMMUNICATIONSLETTERS (2015–2017), and is an Associate Editor of Telecommunications Systems (Springer). He has severed in a large number of international conferences in the capacity of General Co-Chair, TPC Co-Chair, Symposium Chair, and so on.