MTSA-Net: a multiscale time self-attention network for ship radiated self-noise reduction

Hailun Chu, Chao Li, Haibin Wang, Jun Wang, Yupeng Tai, Lei Zhou, Fan Yang, Yannick Benezeth

Highlights

MTSA-Net: a multiscale time self-attention network for ship radiated self-noise reduction

Hailun Chu,Chao Li,Haibin Wang,Jun Wang,Yupeng Tai,Yonglin Zhang,Lei Zhou,Fan Yang,Yannick Benezeth

- This study reformulates the issue of reducing ship radiated self-noise as a signal separation task involving two types of signals: the desired signal from a cooperative source and the self-noise from the towed ship.
- This paper introduces a novel approach, the multiscale time self-attention network (MTSA-Net), which is designed based on the cooperative source structure.
- The proposed model MTSA-Net can obtain a high coefficient gain when maintaining sufficient discrimination.

MTSA-Net: a multiscale time self-attention network for ship radiated self-noise reduction

Hailun Chu[a,b,c,1], Chao Li[a,b], Haibin Wang[a,b,*], Jun Wang[a,b], Yupeng Tai[a,b], Yonglin Zhang[a,b], Lei Zhou[a,b,c], Fan Yang[d] and Yannick Benezeth[c]

[a]State Key Laboratory of Acoustics, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China
[b]University of Chinese Academy of Sciences, Beijing 101408, China
[c]Laboratory ImViA, Universite Bourgogne Franche-Comte, Dijon 21078, France
[d]Laboratoire d'Etude de l'Apprentissage et du Développement – UMR CNRS 5022, Dijon 21078, France

ARTICLE INFO

Keywords:
Underwater signal detection
Ship radiated self-noise
Deep learning
Self-attention
Low signal-to-noise ratio

ABSTRACT

For a surface ship carrying a towed array, its in-band radiated self-noise is one of the near-field strong interference, which will seriously limit the performance of the underwater acoustic (UWA) signal detection system. Typically, the conventional ship noise cancellation methods requires time-synchronized hydrophone arrays, such as the towed linear array (TLA), to suppress its self-noise by using spatial filters. However, the spatial filter based methods fail when the direction of arrival of the desired signal is in the ship noise masking area. In cooperative scenarios, the prior knowledge of the template signal provides additional temporal information, which can be utilized to design a time-domain representations based detection system. In this paper, a multiscale time self-attention network (MTSA-Net) is proposed to mitigate the ship radiated self-noise and enhance the desired signal to improve the performance of signal detection system. Experimental results based on sea trial data indicate the effectiveness of our proposed method.

1. Introduction

For the underwater acoustic (UWA) detection platform with a towed ship as the carrier, its self-noise is a kind of near-field significant interference. The characteristics of the interference radiated by the towed ship primarily depend on the ship type and its mode of operation, encompassing both line spectrum and continuous spectrum components (Arveson and Vendittis, 2000; Carey et al., 1997; Han et al., 2020). Due to the multipath effect, the masked area of the jamming exhibits the phenomenon of angular spread close to the end-fire direction (Feng et al., 2018; Hui et al., 2018; Liang et al., 2023). In addition to developing noise reduction technology at the physical layer (Smith and Rigby, 2022), a highly effective detection algorithm becomes an indispensable requirement in the presence of ship self-noise.

To suppress the self-noise, a range of spatial filter based noise suppression methods have been developed in passive sonar. The adaptive beamformings, including null-steering beamforming (NSBF) (Gershman et al., 1995), minimum variance distortionless response (MVDR) (Capon, 1969), aim to form nulls in the direction of interference. NSBF projects the array data onto the self-noise orthogonal subspace, thereby forming a null-steering beam, with the assumption that the interference direction is known. Because the masking area of the ship self-noise is too large, a large detection blind area will be formed. MVDR maximizes the signal-to-noise ratio by minimizing the noise variance while preserving the desired signal, resulting in enhanced signal clarity and improved detection performance. This method may exhibit diminished performance when the covariance matrix is poorly estimated due to factors like low signal-to-noise ratio (SNR), limited snapshots, and array phase mismatch (Ijsselmuide and Beerens, 2001). There are some robust algorithms that can reduce the mismatch phenomenon at the expense of some high-resolution performance (Yang et al., 2018; Yan and Ma, 2005; Shi et al., 2019). These methods still exhibit a limited impact on angle-spread noise.

Besides, noise suppression methods relying on self-noise estimation are aligned with the inherent background noise of the ship (Liang et al., 2023). These methods comprise three main steps: self-noise bearing estimation, self-noise beam reconstruction, and self-noise cancellation. The typical algorithms are postbeamformer interference canceller (PIC), element interference canceller (EIC), and inverse beamforming (IBF). For the above methods, conventional beamforming (CBF) can be employed to estimate the direction of arrival (DOA) of self-noise (Van Veen and Buckley, 1988), and the beam output from this direction is utilized as a reference signal. The PIC method utilizes an adaptive filter in the beam domain for interference cancellation (Godara, 1991), resulting in a beam pattern with narrow nulls and high side lobes. The EIC method, on the other hand, employs an adaptive filter in the element domain for interference cancellation (Chi et al., 2020). Finally, the IBF method performs interference cancellation in the array element domain, assuming a plane wave scenario (Wilson et al., 2006). Besides, the reference signal can also be acquired by a sensor close to the towed linear array (TLA). The adaptive cancellation methods often exhibit extended convergence times and may encounter challenges in effectively suppressing non-stationary noise originating from the tow ship. While the IBF method

*Corresponding author
chuhailun19@mails.ucas.ac.cn (H. Chu); whb@mail.ioa.ac.cn (H. Wang)
ORCID(s): 0000-0003-3833-2521 (H. Chu)
[1]This is the first author footnote.

exhibits stronger robustness, its assumption of far-field plane waves can introduce array phase mismatch when applied to near-field noise.

The conventional self-noise cancellation methods obtain the ship reference signal by using spatial filters without considering cooperative scenarios. In practice, these methods prove ineffective in scenarios where the desired signal and the noise arrive from the same direction or in non-array application contexts. In these cases, no features in the spatial domain are accessible for signal detection, and there is no enhancement in SNR. In cooperative scenarios, the prior knowledge of cooperative source provides the additional temporal information, which can be used to further improve the detection performance. The matched filter (MF) is an optimal linear filter designed to get time-bandwidth product gain (Turin, 1960), which is solely associated with the time-bandwidth product. However, the gain is compromised by non-Gaussian and non-stationary noise.

This study reformulates the issue of reducing ship radiated self-noise as a signal separation task involving two types of signals: the desired signal from a cooperative source and the self-noise from the towed ship. By leveraging historical data, deep learning (DL) based methods can overcome the limitations associated with conventional spatial filter based methods to acquire knowledge of the ship radiated self-noise. The prevalent DL-based denoising methods can be classified into two categories: the time-frequency (T-F) domain methods (Hao et al., 2021; Lv et al., 2021; Zheng et al., 2021) and the time-domain methods (Luo et al., 2020; Luo and Mesgarani, 2019; Pandey and Wang, 2019). In general, the former is simpler to explain, while the latter exhibits superior performance (Luo and Mesgarani, 2019). The time domain denoising methods utilize a learnable convolutional layer to replace the short time Fourier transform (STFT), which is adopted as the mainstream (Chen et al., 2020; Lam et al., 2021; Subakan et al., 2021). The effectiveness of time-domain denoising methods has been proven in UWA signal detection system. For example, a modified fully-convolutional time-domain audio separation network (ConvTasNet) is created to alleviate ambient noise from the cooperative source (Chu et al., 2023). This network integrates a stacked auto-encoder with a temporal convolutional network for enhanced performance. Additionally, the dual-path transformer network (DPTN) is introduced to reduce ambient noise mixed with ship radiated noise, which combines the multi-head attention transformer and dual-path framework processing (Song et al., 2022). However, in low SNR sceanarios, the aforementioned methods face challenges in capturing long-term dependencies effectively. Moreover, it's crucial to acknowledge that while previous models can learn noise distributions, they often overlook the significance of cooperative signal structures, neglecting to incorporate them into the model's design.

To address the above problems, a multiscale time self-attention network (MTSA-Net) is proposed to denoise the ship radiated self-noise in the cooperative signal. The network consists of three modules: an encoder, a separator, and a decoder. The encoder and decoder are one-layer convolutional blocks, aiming to generate a representation of time-domain signal and transform the representation back to the waveform, respectively. Rather than adopt the sine and cosine functions as the orthogonal basis like STFT (Mitra, 2001), the encoder has a learnable convolutional kernel expected to produce high-level representations. Furthermore, the separator is constructed to form a masking map, which employs a multiscale time self-attention mechanism that simultaneously performs a low-cost self-attention over the full sequence and another dual-path self-attention on local chunks. The sequence structure of multiple time scales is illustrated in Fig. 1. In the context of multiple pulses, the entire sequence is treated as the overarching global entity. Signals with this structure not only share the same time-bandwidth product but also exhibit inter-pulse periodicity. Within this context, we distinguish between intra-pulse and inter-pulse segments, categorizing them as the fast-time and slow-time components, respectively. To facilitate this distinction effectively, our architecture incorporates two attention modules. The global attention module is specifically crafted to capture the overarching characteristics of the sequence, while the dual-path local attention module is designed to discern both intra-pulse intricacies and inter-pulse periodic features. In particular, inter-pulse periodic features result from the multi-pulse structure of the cooperative source, which are not available for non-cooperative source scenes. This combined attention mechanism empowers MTSA-Net with a holistic, global interaction capability. As a result, MTSA-Net demonstrates superior performance in UWA environments compared to previous models.

This paper is organized as follows. In Section 2, the conventional interference cancellation methods are introduced. The MTSA-Net is proposed in 3. Section 4 examines the performance of MTSA-Net. Finally, Section 5 concludes the paper.

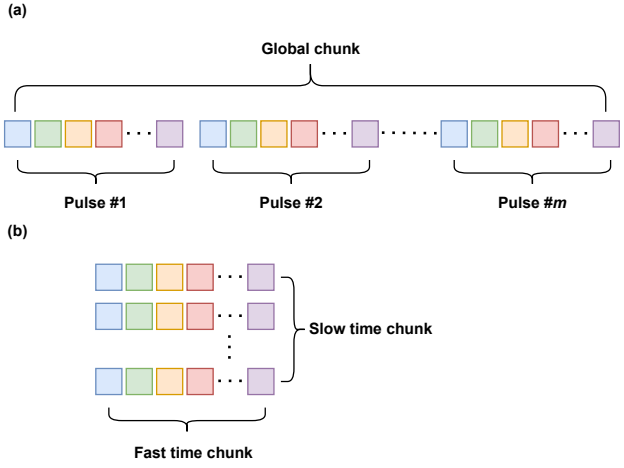

Figure 1: Schematic diagram of multiple time scales of one sequence: (a) The sequence arranged in global time scale; (b) The sequence reorganized in slow and fast time scales.

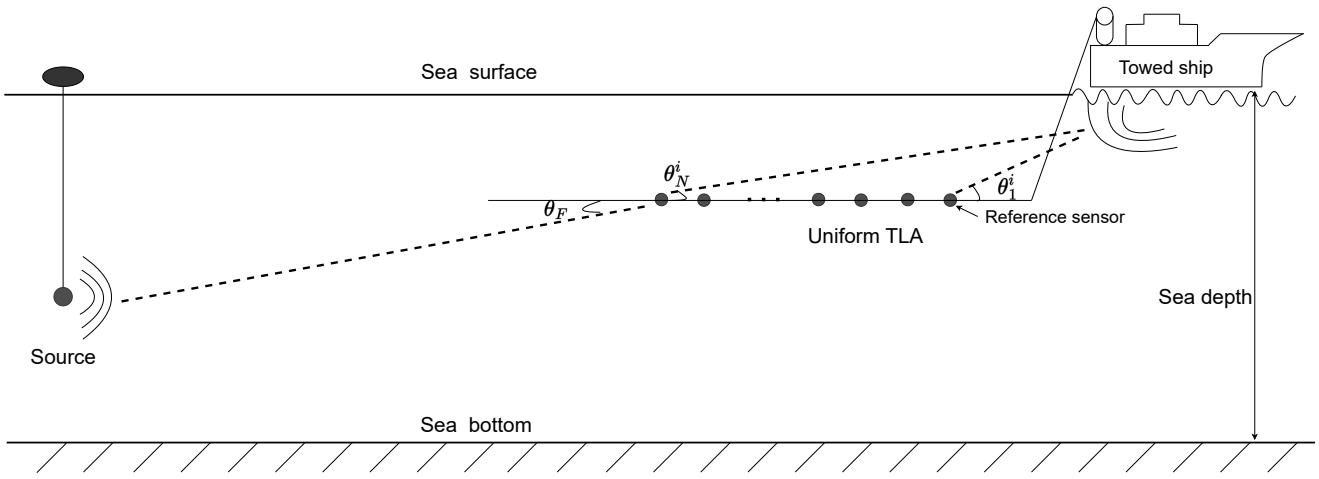

Figure 2: Schematic of the far-field signal reception under the background of towed ship self-noise.

2. Conventional noise cancellation methods

This section introduces several representative conventional noise cancellation methods, along with the signal model and the MF. The ensuing discussion addresses the limitations of these methods.

2.1. Signal model

There exists a far-field cooperative point source emitting well-defined broadband signals, and one nearby towed ship with a uniform spaced TLA that acts as self-noise. The schematic diagram of this scenario is depicted in Fig. 2. The ship length and TLA length are within the range of tens of meters, while the cooperative source is several kilometers or even dozens of kilometers away.

The ship self-noise is mainly composed of propeller noise, structure-borne noise and so on. For instance, the propeller and engine may locate at the stern and middle of the ship, which are considered as different near-field point sources. Therefore, the cooperative source is modeled as a planar wave signal, which can be reached at the TLA from an azimuth θ_F ranging from $[0, 2\pi]$. The ship self-noise is characterized as multiple spherical wave signals. The i-th source approaches the n-th sensor from a DOA of θ_n^i. Let I be the number of near-field sources and assume that the TLA has N sensors with element spacing d. Then, the received signal at frequency f of n-th sensor can be expressed as

$$X_n(f) = S(f)g_n(f, r_F) + \sum_{i=1}^{I} S_i(f)g_n^i(f, r_{N_i}) + \eta_n(f) \quad (1)$$

where $S(f)$ is the cooperative source spectrum, $S_i(f)$ is the i-th towed ship near-field source spectrum and $\eta_n(f)$ is the ambient noise spectrum of n-th sensor. Besides, $g_n(f, r_F)$ and $g_n^i(f, r_{N_i})$ are the Green's functions of n-th sensor at position r_F and r_{N_i}, respectively. The array received data is denoted as $\boldsymbol{X}(f) = [X_1(f), X_2(f), \dots, X_N(f)]^H$.

In this study, the cooperative source adopts the chaotic frequency modulation (CFM) (Shu et al., 2016) based multi-pulse signal, where the i-th pulse has a formulation as

$$s_{i,l}(t) = Re[a(t)e^{j2\pi f_l t}e^{j2\pi f_c t}], (i-1)T_0 \le t < iT_0 \quad (2)$$

where l is the index of modulating symbols, f_c is the carrier frequency and T_0 is the time length of each pulse. Besides, the $f_l = k_l \Delta f$ is the frequency of l-th symbol. Herein, the $Kent$ map (Liu et al., 2015) is used to represent k_l, which ranges from -1 to 1 and Δf is half of the bandwidth. The number of pulses is denoted as m, which is shown in Fig. 1.

2.2. PIC algorithm

At each frequency bin, the signals are processed by forming two beams. Here, the weights of the CBF at frequency f for beam θ are adopted for both the signal beam and self-noise beam that

$$\boldsymbol{w}(f, \theta) = \frac{1}{\sqrt{N}}[1, e^{j2\pi f dcos\theta/c}, \dots, e^{j2\pi f(N-1)dcos\theta/c}] \quad (3)$$

where c is the reference sound speed. The weights of the signal beam and self-noise beam can be denoted as $\boldsymbol{w}(f, \theta_F)$ and $\boldsymbol{w}(f, \theta_N)$, respectively.

Then, the output of the signal beam and the self-noise beam at frequency f are, respectively, given by

$$Beam(f, \theta_F) = \boldsymbol{w}(f, \theta_F)^H X(f) \quad (4)$$

and

$$Beam(f, \theta_N) = \boldsymbol{w}(f, \theta_N)^H X(f) \quad (5)$$

At each frequency bin, the output of the PIC is formed by subtracting the weighted output of the self-noise beam from the signal beam, thus

$$Beam^{PIC}(f) = Beam(f, \theta_F) - h^{PIC}(f)Beam(f, \theta_N) \quad (6)$$

The total output power can be denoted as

$$P^{PIC} = \sum_{f=f_l}^{f_h} |Beam^{PIC}(f)|^2 \quad (7)$$

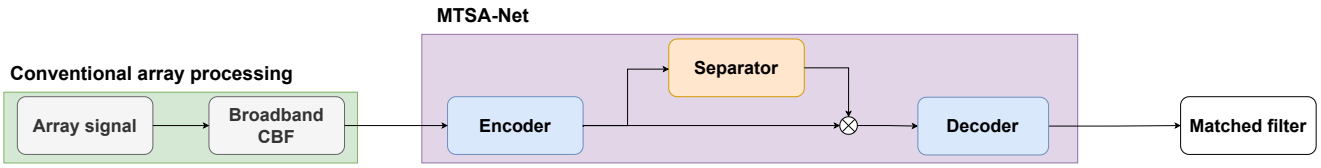

Figure 3: The overall architecture of the MTSA-Net based UWA signal detection system.

where f_l and f_h are the lower and upper frequency limits of the band, respectively. By minimizing the total output power, the optimal weights $\hat{h}^{PIC}(f)$ can be given as

$$\hat{h}^{PIC}(f) = \frac{Beam(f,\theta_F)(Beam(f,\theta_N))^H}{Beam(f,\theta_N)(Beam(f,\theta_N))^H} \tag{8}$$

The PIC is applied to the strongest self-noise beam first and then this procedure can be repeated several times until all self-noise beams present are suppressed.

2.3. EIC and IBF algorithms

The self-noise beam is formed as that of the PIC algorithm. The difference is that the signal beam is formed after canceling self-noise in the element domain.

From Eq. (5), the DOA of strongest self-noise is estimated. The estimated self-noise of the n-th sensor at frequency f can be denoted as

$$\Phi_n(f) = Beam(f,\theta_N)e^{2\pi f(n-1)dcos\theta_N/c} \tag{9}$$

The element domain output of the EIC is formed by subtracting the weighted output of the self-noise beam after phase compensation from the received signal for each sensor,

$$X_n^{EIC}(f) = X_n(f) - h_n^{EIC}(f)\Phi_n(f) \tag{10}$$

the optimal weights $\hat{h}_n^{EIC}(f)$, which minimizes the output power of each sensor, can be given as

$$\hat{h}_n^{EIC}(f) = \frac{X_n(f)\Phi_n(f)^H}{\Phi_n(f)\Phi_n(f)^H} \tag{11}$$

The subtraction process of IBF is performed without adaptive weighting, that is

$$X_n^{IBF}(f) = X_n(f) - \Phi_n(f) \tag{12}$$

After the self-noise cancellation, the signal beam is formed as follows:

$$Beam^{EIC}(f,\theta_F) = \boldsymbol{w}(f,\theta_F)^H X^{EIC}(f) \tag{13}$$

and

$$Beam^{IBF}(f,\theta_F) = \boldsymbol{w}(f,\theta_F)^H X^{IBF}(f) \tag{14}$$

Both algorithms, similar to PIC, can be iteratively repeated multiple times until the self-noise is completely suppressed. However, these methods have limitations when the DOA of the desired signal, i.e. the cooperative source signal, is in the masking area formed by the towed ship self-noise. In this case, the desired signal and ship noise have the same array gain of $10log_{10}N$ after CBF. The output power of the above self-noise cancellation methods will be zero as the signal and interference are canceled together based on the minimum output power criterion. To be pointed out, this limitation is the main issue to be solved. In other cases, the performance experiences degradation due to the self-noise beam having a non-zero response in the signal direction, along with a portion of the signal leaking into the self-noise beam.

2.4. Matched filter

In cooperative scenarios, the cooperative source spectrum $S(f)$ is the prior knowledge. Therefore, the MF can be applied to get additional time-bandwidth product gain. After self-noise cancellation, the signal beam can be denoted as $Beam^{output}(f,\theta_F)$ representing the above three methods. The input of MF in time domain is

$$Beam^{output}(t,\theta_F) = IFFT\{B^{output}(f,\theta_F)\} \tag{15}$$

The impulsive response of MF in time domain is

$$h(t) = s(\tau - t) = IFFT\{S(f)^H e^{j2\pi f\tau}\} \tag{16}$$

Then, the output of MF is obtained as

$$y(t,\theta_F) = Beam^{output}(t,\theta_F) \otimes h(t) \tag{17}$$

where $\otimes$ represents the convolution operation. The theoretical gain is

$$G_{TBP} = 10\log_{10} 2BW \cdot T \tag{18}$$

where BW is the bandwidth and T is the time duration. In order to improve the stability of MF, its normalized form is adopted.

3. The multiscale time self-attention network

The overall architecture of the MTSA-Net based UWA signal detection system is illustrated in Fig. 3. Once the array processing is completed, the beam output is fed into the MTSA-Net for augmentation. The MF is a temporal filter serving as a detector. The loss functions are shown in Section 3.3.

Fig. 4 presents the detailed architecture of the MTSA-Net, which consists of encoder, decoder and separator modules. During the training of MTSA-Net, both the self-noise

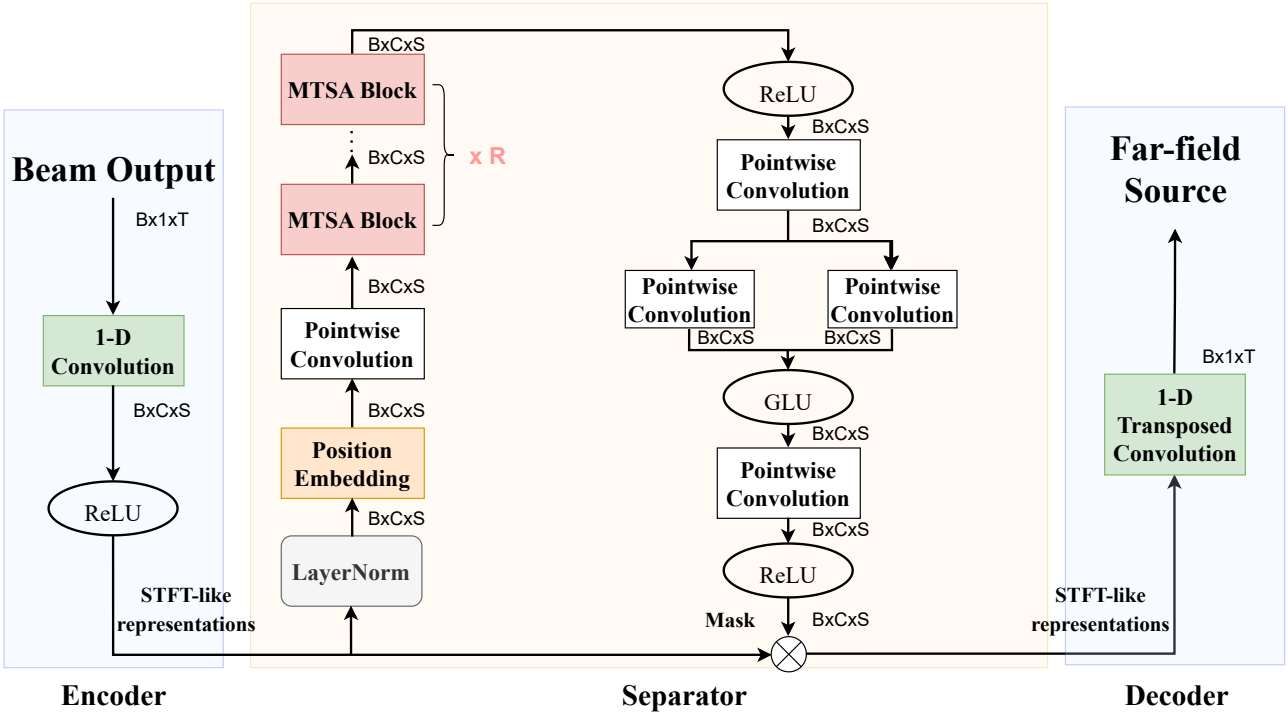

Figure 4: The architecture of the MTSA-Net.

and noisy cooperative signal are taken as input in pairs, while the corresponding generated outputs are the self-noise and pure cooperative signal, respectively. Throughout the testing process, the self-noise and noisy cooperative signal are provided separately as the negative and positive samples.

3.1. Encoder and decoder

The encoder is utilized to extract the STFT-like representations in a latent feature space. It consists of a one-dimensional convolutional layer (Conv1D) and a rectified linear unit (ReLU). The kernel size of Conv1D is L with a stride of $L/2$. The input sequence $\boldsymbol{input} \in R^{B\times 1\times T}$ is encoded to the feature map Z:

$$\boldsymbol{Z} = ReLU(Conv1D(\boldsymbol{input})) \tag{19}$$

where $\boldsymbol{Z} \in R^{B\times C\times S}$, C is number of filters, B is the batch size and $S = \lceil (T - L/2)/(L/2) + 1 \rceil$. The notation $\lceil \cdot \rceil$ is rounding up. Zero padding is used to ensure each channel having the same time length.

Let the output of separator, i.e. the mask, be denoted as $\boldsymbol{M} \in R^{B\times C\times S}$. Then, the masked feature map $\boldsymbol{Z}_M = \boldsymbol{Z} \odot \boldsymbol{M}$ is decoded into waveform $\boldsymbol{output} \in R^{B\times 1\times T}$:

$$\boldsymbol{output} = TransposedConv1D(\boldsymbol{Z}_M) \tag{20}$$

where $\odot$ is the element-wise multiplication operation, and the decoder is one Transposed Conv1D layer with same kernel size and stride as the encoder.

3.2. Separator

The separator is designed to apply a non-linear mapping from the feature map $\boldsymbol{Z}$. To accomplish this, the feature map $\boldsymbol{Z}$ performs layer normalization and is added with sinusoidal positional encodings (Vaswani et al., 2017b) in position embedding block. The position embedding provides positional information for the learning of attention modules. Afterward, the sequence undergoes a pointwise convolution before being input to the MTSA block.

In the MTSA block, the sequence is processed by the attention mechanism. The proposed attention mechanism consists of three scales time self-attention: fast time, slow time and global time. It performs a joint local and global self-attention operations, in which the skip connection is adopted for ease of training. The process of the MTSA block is repreated R times.

After the sequence passes through the ReLU and a pointwise convolutional layer, it is passed to a parallel pointwise convolutional layer and a gate linear unit (GLU) (Shazeer, 2020). Finally, the sequence undergoes another pointwise convolution, followed by a ReLU activation, to yield the mask $\boldsymbol{M}$.

3.2.1. MTSA block

The MTSA block is developed to integrate the global self-attention and local self-attention for feature extraction in three time scales, whose architecture is presented in Fig. 5(a). The processing flow is as follows. First, the feature input $\boldsymbol{F}_{in} \in R^{B\times C\times S}$ is processed by the token shift mod-

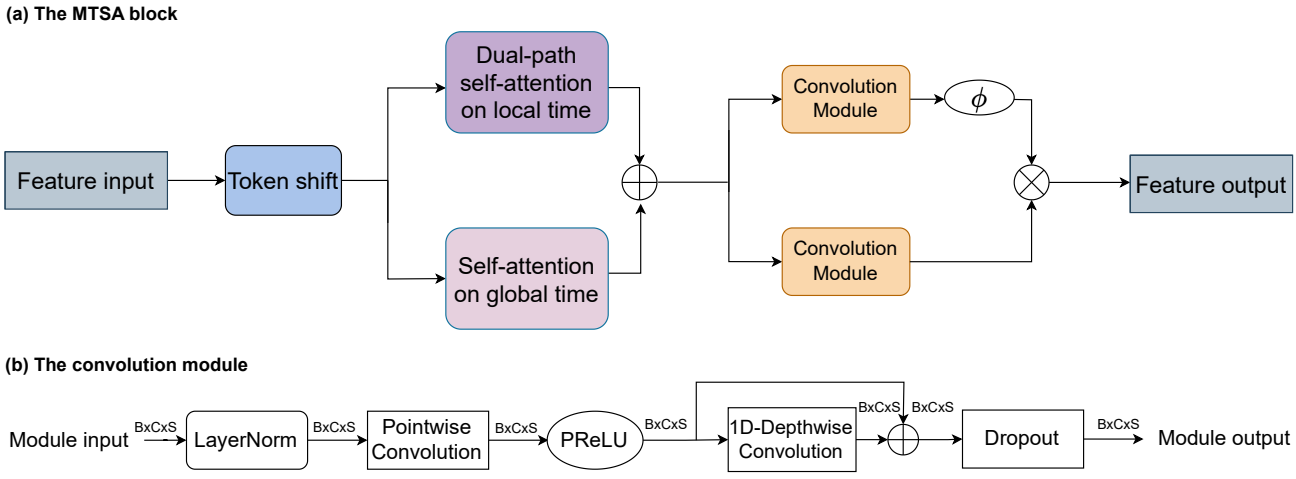

Figure 5: (a) The diagram of the MTSA block. (b) The diagram of the convolution module.

ule, which is a simple offset in the temporal dimension at each block and has almost no computational cost (Peng et al., 2023). It is implemented by shifting the first half of the channels while preserving the last half channels. The shifted channels primarily collecte contextual information from the preceding tokens and pass it to the following tokens, whereas the unshifted channels are primarily responsible for prediction (Peng et al., 2023). Then, the token shifted feature $\boldsymbol{F}_{shift} \in R^{B\times C\times S}$ is fed parallelly into the dual-path local self-attention module (DPL-SA) and the global self-attention module (GSA), getting the augmented features $\boldsymbol{F}_{loc} \in R^{B\times C\times S}$ and $\boldsymbol{F}_{glo} \in R^{B\times C\times S}$. Next, the augmented features are added to obtain the integrated features $\boldsymbol{F}_{aug} \in R^{B\times C\times S}$. Moreover, there is a gating operation applied to the integrated features to improve the capability of the MTSA block. It has the ability to selectively filter input information, extract crucial features, and model long-term dependencies, thereby enhancing the model's expressive and generalization capabilities (Hua et al., 2022; Narang et al., 2021). In this operation, two convolution modules are conducted, which is illustrated in Fig. 5(b). The kernel size of 1D-Depthwise convolution is L_1 and zero-padding is adopted. This module utilizes the combination of pointwise convolution and 1-D-Depthwise convolution to reduce the number of required parameters (Chollet, 2017). Let the convolution module be denoted as function $Conv\text{-}M$. The process of MTSA block is described as

$$\boldsymbol{F}_{shift} = Token\text{-}shift(\boldsymbol{F}_{in}) \tag{21}$$
$$\boldsymbol{F}_{loc} = DPL\text{-}SA(\boldsymbol{F}_{shift}) \tag{22}$$
$$\boldsymbol{F}_{glo} = GSA(\boldsymbol{F}_{shift}) \tag{23}$$
$$\boldsymbol{F}_{aug} = \boldsymbol{F}_{loc} + \boldsymbol{F}_{glo} \tag{24}$$
$$\boldsymbol{F}_{out} = Conv\text{-}M(\boldsymbol{F}_{aug}) \cdot \phi(Conv\text{-}M(\boldsymbol{F}_{aug})), \tag{25}$$

where the $\boldsymbol{F}_{out}$ is the output of MTSA block and $\phi(\cdot)$ is the $Sigmoid$ function.

In dual-path self-attention module as depicted in Fig. 6 (a), the module input $\boldsymbol{F}_{shift}$ is firstly split into non-overlapping chunks of length P. The last chunk is zero-padded to generate H equal size chunks and all chunks are then concatenated together. Then, the long sequence with size S is segmented into inter-chunks with size P and intra-chunks with size H, which represent fast time and slow time scale, respectively. Afterwards, the scaled dot-product self-attention (SDPSA) is applied for the two paths by groups (Vaswani et al., 2017a), which is shown in Fig. 6 (c). The two SDPSA modules can be defined as:

$$\boldsymbol{F}_{loc}^{fast} = (softmax(\boldsymbol{Q}_{loc}^{fast}(\boldsymbol{K}_{loc}^{fast})^T))\boldsymbol{V}_{loc}^{fast} \tag{26}$$
$$\boldsymbol{F}_{loc}^{slow} = (softmax(\boldsymbol{Q}_{loc}^{slow}(\boldsymbol{K}_{loc}^{slow})^T))\boldsymbol{V}_{loc}^{slow} \tag{27}$$

where $\boldsymbol{Q}_{loc}^{fast}$, $\boldsymbol{K}_{loc}^{fast}$, and $\boldsymbol{V}_{loc}^{fast} \in R^{P\times C}$ and $\boldsymbol{Q}_{loc}^{slow}$, $\boldsymbol{K}_{loc}^{slow}$, and $\boldsymbol{V}_{loc}^{slow} \in R^{H\times C}$. The $B \times H$ and $B \times P$ represent the group size in fast and slow time dimensions, respectively.

As depicted in Fig. 6 (b), the global attention module has the same input as the local attention module, where $C < S$. To speed up the computation, a low cost self-attention is shown in Fig. 6 (d) and derived (Zhuoran et al., 2021) as:

$$\boldsymbol{F}_{glo} = \boldsymbol{Q}_{glo}(sigmoid(\boldsymbol{K}_{glo})^T sigmoid(\boldsymbol{V}_{glo})) \tag{28}$$

where $\boldsymbol{Q}_{glo}$, $\boldsymbol{K}_{glo}$, and $\boldsymbol{V}_{glo} \in R^{S\times C}$. The B represents the group size in global attention module.

To explain the complexity of computation, the $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ can be denoted with the dimension $S \times C$. Without the $Softmax$ and $Sigmoid$ function, the multiplication would involve three matrices $\boldsymbol{Q}\boldsymbol{K}^T\boldsymbol{V}$. Considering the matrix multiplication is associative, $\boldsymbol{K}^T\boldsymbol{V}$ can be calculated firstly, resulting in a $C \times C$ matrix, and then left multiply it with $\boldsymbol{Q}$. Since $C < S$, this computation yields an approximate complexity of $\mathcal{O}(S)$, which is dominated by the $\boldsymbol{Q}$ left multiplication step. If the matrix multiplication is calculated like Eq. 26 and 27, QK^T produces a $S\times S$ matrix, which determines that the complexity of this attention is $\mathcal{O}(S^2)$.

3.3. Loss function

The proposed MTSA-Net is designed to provide accurate estimation of both ship self-noise and pure cooperation

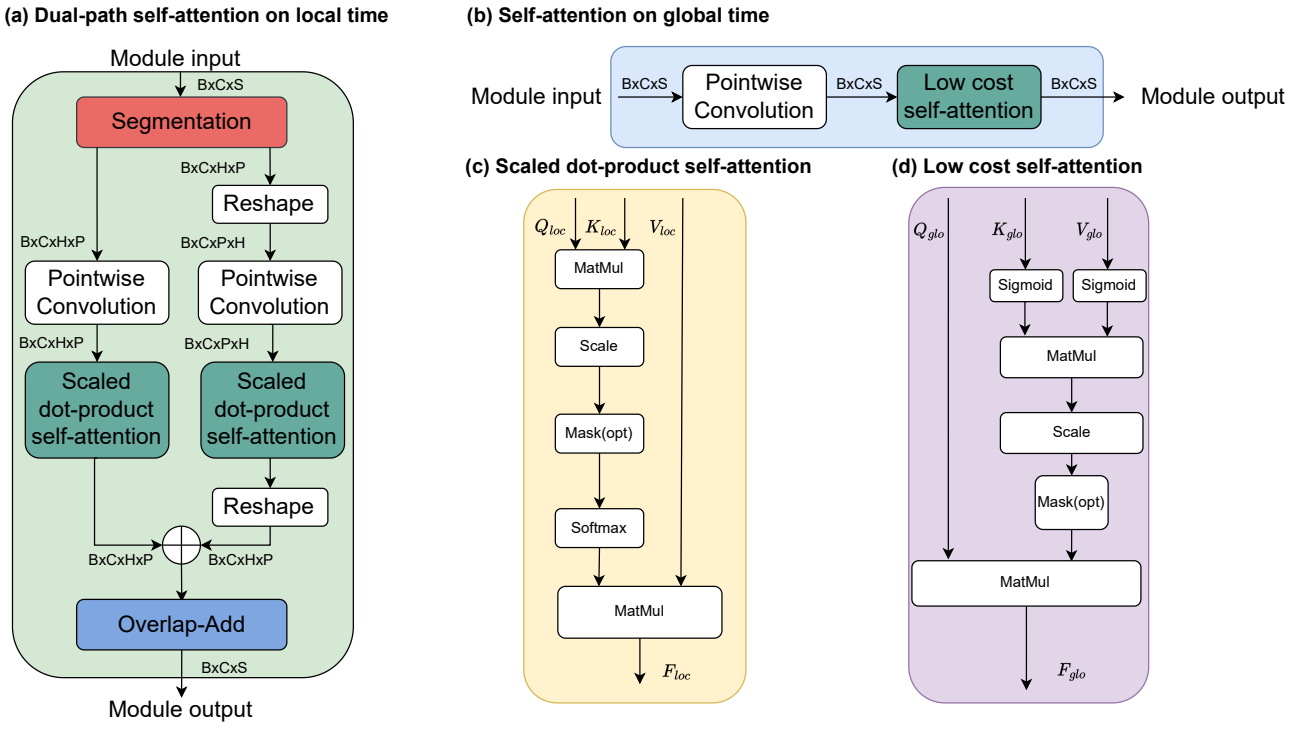

Figure 6: (a) The diagram of the dual-path self-attention on local time module. (b) The diagram of the self-attention on global time module. (c) The detailed structure of the scaled dot-product self-attention. (d) The detailed structure of the low cost self-attention.

source signal. In this section, the modified scale-invariant signal-to-noise ratio (MSI-SNR) is adopted to match the noise information and source information. The MSI-SNR contains the measurements of the original self-noise and source, which is the loss function in this study:

$$Loss_{MSI\text{-}SNR} = SI\text{-}SNR(\boldsymbol{s}, \hat{\boldsymbol{s}}) + SI\text{-}SNR(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}) \quad (29)$$

where $\boldsymbol{s}$ is the target and $\boldsymbol{\eta}$ is the ship noise. The SI-SNR($\boldsymbol{s}$,$\hat{\boldsymbol{s}}$) (Chu et al., 2023) can be calculated as:

$$\begin{cases} \boldsymbol{s}_{proj} = \frac{\langle \hat{\boldsymbol{s}}, \boldsymbol{s} \rangle \boldsymbol{s}}{\|\boldsymbol{s}\|^2} \\ \boldsymbol{e} = \hat{\boldsymbol{s}} - \boldsymbol{s}_{proj} \\ SI\text{-}SNR = 10 \log_{10} \frac{\|\boldsymbol{s}_{proj}\|^2}{\|\boldsymbol{e}\|^2} \end{cases} \quad (30)$$

where $\langle \cdot \rangle$ and $\| \cdot \|$ denote the inner product operation and power operation, respectively. It can be proven that SI-SNR is invariant with the amplitude of signal, which is equivalent to the correlation coefficient (Rugini and Banelli, 2016). In Eq. 29, the first term signifies the capacity to learn from signals, while the second term signifies the ability to learn from noise.

4. Experimental results

In this section, the experimental results are presented, including an introduction of the datasets employed, evaluation metrics, representative baselines, and the evaluation of the performance of the proposed MTSA-Net.

Table 1
The detailed experimental setup.

Cond.	PRM	STW/knots	SOG/knots	AD/m	WD/m	CL/m
Cond1	70	4	3.2	51	144	300
Cond2	80	5.1	4.4	54	138	300
Cond3	90	6	4.9	48	138	300
Cond4	100	6.8	6	46	137.8	320
Cond5	110	7.7	6.8	44	134	320

4.1. Experimental setup and datasets

The raw data used for this study was collected in the shallow water of the South China Sea. While the towed ship was sailing at different revolutions per minute (RPM), the 30-element TLA with a spacing of 1.5 m records its own noise. Additionally, there are no other nearby ships present during the experiment. The sampling frequency f_s is 10 kHz. The detailed experimental setup is shown in Table 1, including the RPM, speed through water (STW), speed over ground (SOG), array depth (AD), water depth (WD) and cable length (CL).

In this experiment, a band-pass filter is applied. For each scan azimuth of broadband CBF, 2508 self-noise (negative) samples are generated, each with a duration of 4.2 seconds. In addition, the noisy signal (positive) samples are generated by adding the target signal in the corresponding negative samples, resulting the same 2508 samples. The data from each of the four conditions forms the training set, while

the remaining condition comprises the test set. Besides, the cross validation is chosen and the ratio of the training set to testing set is approximately 4:1. This allows the noise of the training set and the test set to be collected at different PRM.

Within the training set, the arrival time of the target signal is randomly distributed between [0, 0.2] seconds, and the SNR of each sample varies randomly between [-25, 0] dB. In the testing set, the SNR of each sample ranges from -35 dB to 0 dB. The SNR is defined as:

$$SNR = 10\log_{10}\frac{\|\boldsymbol{s}\|^2}{\|\boldsymbol{\eta}\|^2} \tag{31}$$

To avoid confusions, the near-field interference is the towed ship self-noise. In the following experiments, the denotation of SNR is adopted.

4.2. Evaluation metrics

To assess the performance of the proposed MTSA-Net, the SI-SNR, two DSI-SNRs (Chu et al., 2023), the probability of detection (P_D), the probability of false alarm (P_F) are selected as evaluation metrics.

The two DSI-SNRs are, respectively, expressed as

$$\begin{aligned} DSI\text{-}SNR_1 = &\frac{1}{N_{cv}}\sum_{i=0}^{N_{cv}}(SI\text{-}SNR(\hat{\boldsymbol{s}}_{i|\boldsymbol{H_1}}, \boldsymbol{s})) \\ &- \frac{1}{N_{cv}}\sum_{i=0}^{N_{cv}}(SI\text{-}SNR(\hat{\boldsymbol{s}}_{i|\boldsymbol{H_0}}, \boldsymbol{s})) \end{aligned} \tag{32}$$

$$\begin{aligned} DSI\text{-}SNR_2 = &\frac{1}{N_{cv}}\sum_{i=0}^{N_{cv}}(SI\text{-}SNR(\hat{\boldsymbol{s}}_{i|\boldsymbol{H_1}}, \boldsymbol{s})) \\ &- max(SI\text{-}SNR(\hat{\boldsymbol{s}}_{i|\boldsymbol{H_0}}, \boldsymbol{s})) \end{aligned}, \tag{33}$$

where $\hat{\boldsymbol{s}}_{i|\boldsymbol{H_1}}$ and $\hat{\boldsymbol{s}}_{i|\boldsymbol{H_0}}$ represent model outputs when the i-th positive sample and i-th negative sample are inputted, respectively. N_{cv} is the number of positive samples in the test dataset. Both DSI-SNR$_1$ and DSI-SNR$_2$ quantify the enhancement in SI-SNR. DSI-SNR$_1$ evaluates the average improvement, while DSI-SNR$_2$ specifically examines the enhancement at low P_F.

The P_D and P_F are, respectively, formed as

$$P_D = N_{TT}/(N_{TT} + N_{FT}) \tag{34}$$

$$P_F = N_{TF}/(N_{TF} + N_{FF}), \tag{35}$$

where N_{TT} and N_{FT} are the detected and undetected positive samples, respectively. N_{TF} and N_{FF} are the detected and undetected negative samples, respectively.

4.3. Competitive baselines

In this study, the typical traditional and DL-based methods are investigated to quantitatively compare them with the proposed network. The conventional noise cancellation methods introduced in Sec. 2 are invalid when the DOA of the desired signal is in the ship self-noise masking area. Therefore, the MF is adopted as conventional baseline model. For the DL-based comparison, the ConvTasNet (Luo and Mesgarani, 2019), and DPTN (Chen et al., 2020) are utilized. ConvTasNet is a fully convolutional neural network (CNN), whereas DPTN combines the CNN, recurrent neural network (RNN) and attention mechanism.

Both structures share the same encoder and decoder design with MTSA-Net, differing only in the kernel size. The main difference is in the separator module. ConvTasNet utilizes the temporal convolutional network (TCN) with an increasing dilation factors as its separator. The dilation factors grow exponentially to guarantee an adequately expansive temporal context window, enabling the utilization of the signal's long-range dependencies. In the separator, DPTN splits this module input into overlapped segments. Then, a module combining RNN and multi-head attention combined is designed to learn the order information of the sequence without positional encodings. However, it's worth noting that the segmentation stage does not account for the structural aspects of the source signal and this structure only provides a fast and slow time scale information to approach the global context awareness.

All the above models are retrained on the same datasets introduced in Sec. 4.1. The proposed MTSA-Net is implemented on the advanced DL framework $PyTorch$ by using NVIDIA RTX 4090 GPU. Table 2 compares the model size and computational complexity of above DL-based methods, where the model size of the proposed MTSA-Net isbetween the convTasNet and DPTN models. The five performance metrics mentioned in Sec. 4.2 are used to analyze the experimental results.

The results are presented in Table 3 at three different SNRs, where SI-SNR describes the enhancement effect of signal components and P_D is calculated when P_F is at 10^{-3}.

Table 2
Comparisons of model size and computational complexity.

Model	ConvTasNet	DPTN	**MTSA-Net**
No. Parameters	4.92M	2.64M	3.98M
FLOPs	13.46G	220.76G	469.84M

Table 3
Quantitative comparison between baseline methods and the proposed method.

SNR (dB)	Method	P_D (%)	SI-SNR (dB)	DSI-SNR$_1$ (dB)	DSI-SNR$_2$ (dB)
-20	MF	7.6	-18.24	1.39	-2.53
	ConvTasNet	20.0	-16.58	2.19	-1.76
	DPTN	47.8	-13.98	4.17	-0.20
	MTSA-Net	**62.7**	**5.65**	**23.95**	**18.64**
-18	MF	21.3	-17.22	2.41	-1.51
	ConvTasNet	42.2	-15.22	3.55	-0.40
	DPTN	74.0	-12.90	5.25	0.88
	MTSA-Net	**83.5**	**17.39**	**35.69**	**30.38**
-15	MF	72.2	-14.83	4.79	0.88
	ConvTasNet	87.6	-12.04	6.73	2.78
	DPTN	98.0	-11.04	7.12	2.75
	MTSA-Net	**99.3**	**30.66**	**48.97**	**43.66**

Table 4
Ablation studies on the MTSA block when SNR=-20 dB.

Global time attn.	Fast time attn.	Slow time attn.	P_D (%)	SI-SNR (dB)	DSI-SNR$_1$ (dB)	DSI-SNR$_2$ (dB)
×	✓	✓	43.6	-3.48	14.54	6.97
✓	✓	×	43.8	-2.74	15.99	8.09
✓	×	✓	54.7	3.56	22.47	16.63
✓	✓	✓	**62.7**	**5.65**	**23.95**	**18.64**

It is evident that the conventional MF method exhibits subpar performance in detecting UWA signals with low SNRs due to its susceptibility to noise mismatch. Furthermore, the DL-based baselines consistently outshine the conventional method across all evaluation metrics. Additionally, the proposed MTSA-Net surpasses other DL-based baselines on the aforementioned datasets, across various SNRs, and attains the highest evaluation scores. Taking SNR=-20 dB as an example, it yields a notable 55.1% improvement in P_D, 23.89 dB gain in SI-SNR, 22.56 dB gain in DSI-SNR1 and 21.17 dB gain in DSI-SNR2, which indicates its effectiveness even at low SNRs. These results demonstrate that MTSA-Net is an efficient solution for detecting UWA signals against ship noise background, particularly in low SNR conditions. This superiority arises from their ability to learn and adapt to the features of ship noise and cooperative source signals. More specifically, obtaining high detection probability requires discrimination between positive and negative samples. It can be seen that the baselines attain a relatively high value for P_D, yet other performance metrics do not exhibit equally impressive results. Meanwhile, the proposed MTSA-Net achieves promising results in all metrics. This is attributed to the superior capacity of MTSA-Net to effectively learn the intricate signal structure while concurrently preserving its ability to discern noise characteristics.

4.4. Ablation studies

This section conducts ablation studies on the MTSA-Net and the results are shown in Table 4. The experimental configuration for all models remains consistent with that of MTSA-Net, with the exception being the model architecture. Scores of P_D, SI-SNR, DSI-SNR1, DSI-SNR2 are adopted.

The MTSA block consists of attention mechanisms operating across three distinct time scales: global time, fast time and slow time. To assess the effectiveness of these attention mechanisms, a stepwise approach is employed, systematically removing these mechanisms individually for evaluation. First, the global time attention module in every MTSA block is removed. It means that only the dual-path local time attention module is adopted. In this case, the model achieves the P_D of 43.6%, SI-SNR of -3.48 dB, DSI-SNR1 of 14.54 dB and DSI-SNR$_2$ of 6.97 dB. Besides, the score of P_D is similar to that of DPTN (see Table 3) but the scores of other metrics are improved about 10 dB. While DPTN also incorporates two local time scales, it adopts a sequential learning approach. In this model, the features of these two time scales are learned in parallel and feature fusion is carried out within each module, which results in a better learning of signal structure.

Second, the performance is evaluated by using the global and fast time attention modules. It exhibits performance parity with the preceding model, implying a congruent role for the slow time attention module when compared to the global time attention module. The results show that the two models can integrate contextual information appropriately.

Third, keep the other two attention modules of larger time scales. The performance improves compared to the first two models, indicating the contribution of slow time and global modules are relatively substantial. In summary, the MTSA-Net proposed in this study proficiently extracts both self-noise and signal features from multiple time scales. Notably, when integrated with the fast time attention module, the slow time attention module offers an additional dimension of global awareness, supplementing the global module's capabilities.

Table 5
Summary of model parameters.

Model parameter	B	T	C	L	S	P	R
Value	4	44000	256	50	1760	400	8
Signal parameter	m	T_0 (s)	f_s (Hz)	f_l (Hz)	f_c (Hz)	f_h (Hz)	/ /
Value	4	1	10000	300	350	400	/

Table 6
Parameter optimization on the MTSA block.

SNR (dB)	Method	P_D (%)	SI-SNR (dB)	DSI-SNR$_1$ (dB)	DSI-SNR$_2$ (dB)
-20	P=256	33.8	-8.96	10.58	6.63
	P=300	45.8	-0.77	18.78	14.88
	P=400	**62.7**	**5.65**	**23.95**	**18.64**
	P=450	47.1	-0.31	19.22	15.33

4.5. Parameter optimization

The optimation method is Adam (Kingma and Ba, 2014), learning rate is 0.001, and maximum epoch is 200. To avoid overfitting, the early stopping strategy is adopted. The parameters of MTSA-Net and the cooperative source signal are shown in Table 5. In cooperative sceanarios, the model parameter selection is supposed to be linked to the structure of cooperative signal. In this study, the signal is composed of 4 periodic pulses, spanning a total of 40000 samples. Following the application of an encoder module with a kernel size of 50 and a stride size of 25, a total of 1600 remaining time steps are obtained. Consequently, each individual pulse consists of 400 time steps, equivalent to the size of an intra-chunk.

In order to illustrate the connection between model parameter selection and signal structure, the tuning experiment on intra-chunk size, which is denoted as P, is conducted. The results are shown in Table 6. It is obvious that the model shows the best result when the intra-chunk size is equal to

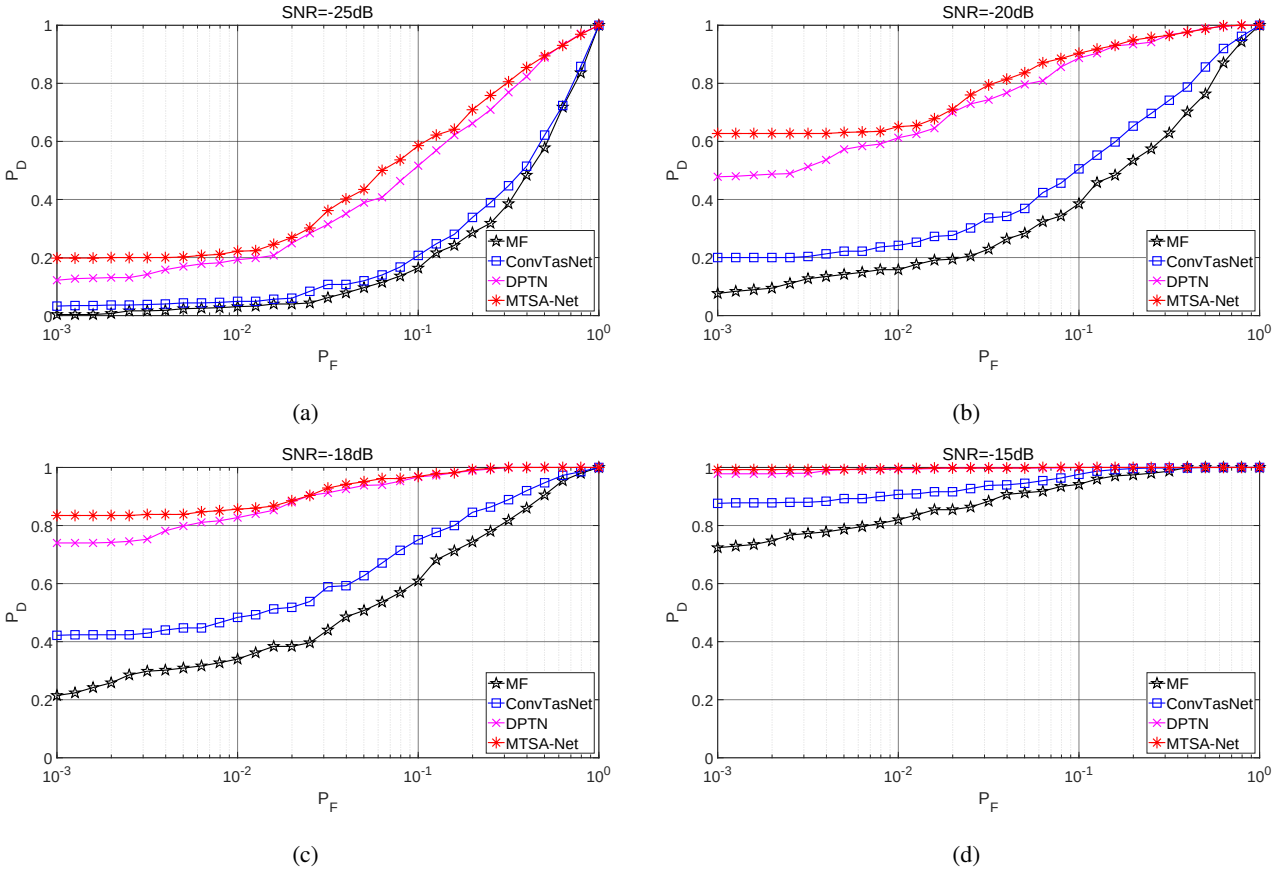

Figure 7: ROC comparisons for different methods in different SNRs: (a) SNR=-25 dB; (b) SNR=-20 dB; (c) SNR=-18 dB; (d) SNR=-15 dB.

400. The performance drops significantly as the intra-block size deviates from 400. When the intra-chunk size is set to either 300 or 450, there is a 15% reduction in P_D, and the other three metrics have a 5 dB decrease. The tuning experiment demonstrates that model with parameters that match the signal structure yields superior performance.

4.6. Detailed statistical analysis

In this section, the ROC curves in various SNRs and SNR versus P_D curves for different P_F are provided to analyze the statistical properties. Besides, the correlation coefficients output by the baselines and MTSA-Net are also shown, facilitating a comprehensive evaluation of the algorithms' performance and its impact on signal enhancement.

First, these ROC curves stand as fundamental graphical tools, shedding light on the intricate dynamics between P_D and P_F, thereby providing valuable insights into the effectiveness of the detection process. The ROC curves of different methods are shown in Fig. 7. It can be seen that the conventional method MF, compared to the DL-based methods, has poor performance in low SNRs, and achieves the P_D of 7.6% when SNR=-20 dB. The performance degradation can be attributed to the non-Gaussian nature of ship noise, resulting in a discrepancy between the noise model employed by MF and the actual noise characteristics. Furthermore, the

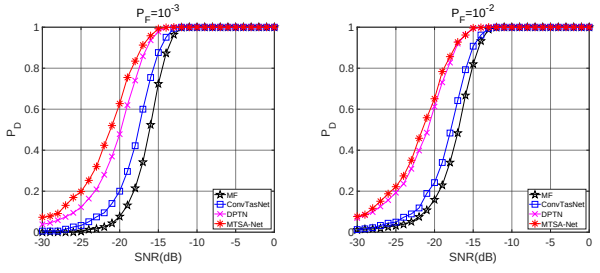

Figure 8: Dependency of PD versus SNR in CFAR detection: (a) $P_F = 10^{-3}$; (b) $P_F = 10^{-2}$.

proposed MTSA-Net is the best model across all baselines for various SNRs, which implies that signal and noise features are well learned and good discrimination is achieved. Moreover, the P_D improvement is the greatest in the low P_F region. This is noteworthy because detection systems exhibiting high P_F values can pose considerable challenges for sonar operators.

Second, the dependence of P_D versus SNR at low P_F holds greater significance, which is investigated in Fig. 8. It is evident that the proposed method outperforms the other

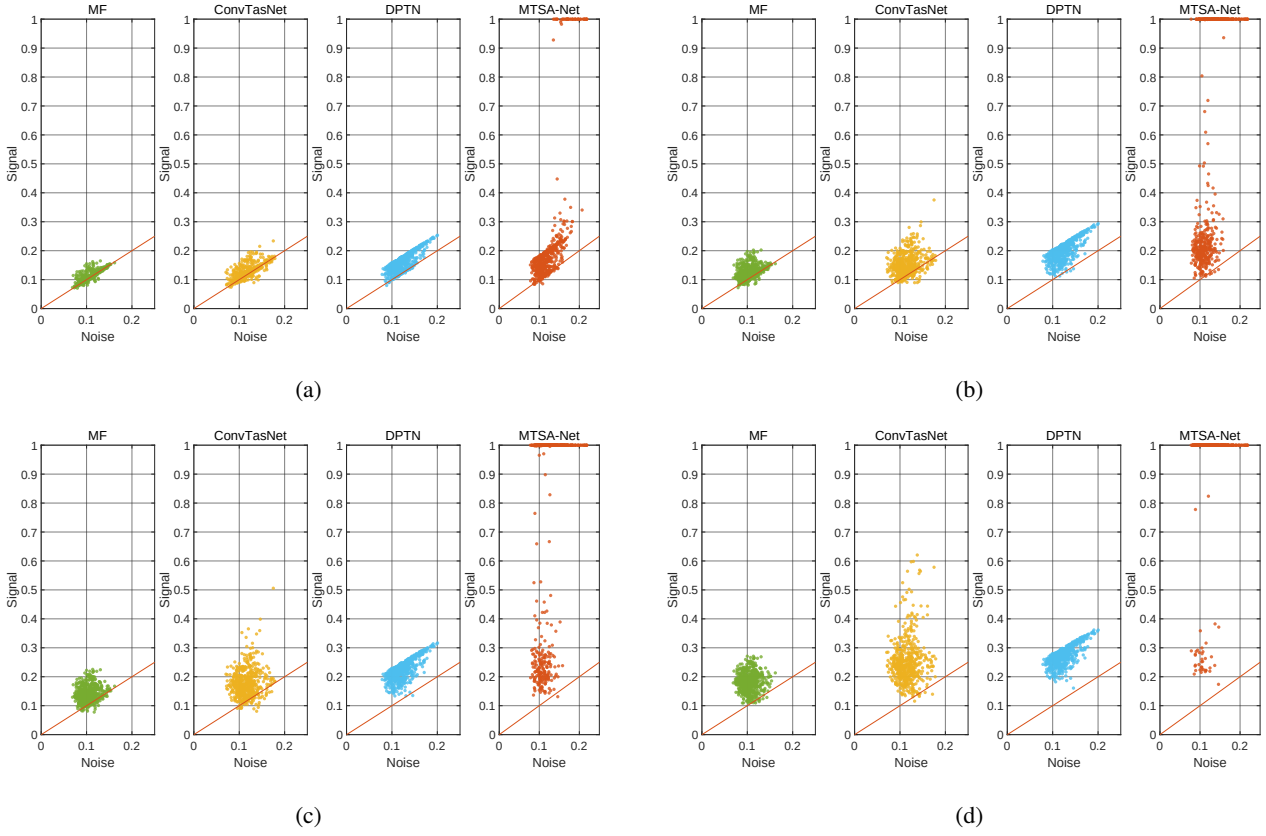

Figure 9: The output correlation coefficients of MF, ConvTasNet, DPTN and MTSA-Net in different SNRs: (a) SNR=-25 dB; (b) SNR=-20 dB; (c) SNR=-18dB; (d) SNR=-15 dB.

baselines consistently across various SNRs. When $P_F = 10^{-3}$, The SNR gain of the MTSA-Net relative to the baselines is 1.5 dB, 4 dB, and 5.5 dB, respectively. Upon comparing the outcomes concerning P_F under different conditions, it is apparent that the MTSA-Net introduced in this study yields greater advantages, particularly in the context of low P_F. This further proves the effectiveness of the proposed method.

In addition to the evaluation metrics of P_D and P_F, the correlation coefficient is a normalized indicator that can evaluate the quality of the signal. Through a meticulous comparison of correlation coefficients among positive and negative samples, with the replicated signal, before and after undergoing various model enhancements, a more lucid assessment of the performance enhancements offered by different models across the entire datasets becomes discernible and the results are shown in Fig. 9. It is apparent that the output correlation coefficients of MF are gathered in lower left corner and the correlation coefficients increase slowly with SNR. For ConvTasNet, the coefficient gain exhibits a significant increase in only a subset of samples. In terms of DPTN, the discrimination between positive and negative samples is extremely pronounced but the coefficient gain is limited. This observation provides insight into the condition wherein a high detection rate is achieved without a corresponding significant gain in SI-SNR, DSI-SNR1 and DSI-SNR2. The

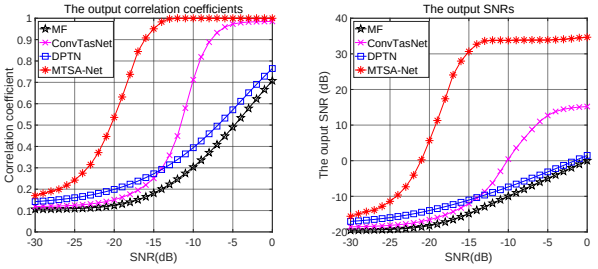

Figure 10: (a) Dependency of mean output correlation coefficients versus input SNRs; (b) Dependency of mean output SNRs versus input SNRs;

proposed MTSA-Net can still obtain a high coefficient gain when maintaining sufficient discrimination. Additionally, the gain of certain samples can attain extremely high levels. It implies that the learning of noise characteristics is good enough to guarantee low P_F and the learning of signal characteristics is significantly better than other baselines. Furthermore, once the signal characteristics are identified, the likelihood of the output correlation coefficient value being close to 1 is remarkably high, consequently leading to significantly improved evaluation metrics.

To further illustrate the enhancement capability of the

MTSA-Net, the mean output correlation coefficients and output SNRs versus input SNRs are depicted in Fig. 10a and 10b, respectively. The growth trend of MTSA-Net and ConvTasNet exhibit similarity, as do those of MF and DPTN. When the correlation coefficients among the three baselines reach approximately 0.3, the corresponding value for MTSA-Net exceeds 0.9. As for ouput SNR, when the values among the three baselines are around -10 dB, the corresponding value for MTSA-Net rise above 30 dB. For MF and DPTN, the increase of output correlation coefficients and SNR is slow. Besides, the output correlation coefficient and SNR of MTSA-Net become stable when input SNR is above -15 dB. In this circumstance, the coefficients of most samples are equal to 1, representing that the positive samples are denoised and successfully enhanced.

5. Conclusion

In this study, a DL-based MTSA-Net is proposed for denoising the ship radiated self-noise under cooperative source scenario. The network comprises three modules, in which the encoder and decoder module generate a representation of time-domain signal and transform the representation back to the waveform, respectively. Besides, the separator consisting of multiple multiscale time self-attention blocks is designed to integrate the local information of intra-pulse and inter-pulse with global information. Especially, the learning of inter-pulse features is based on the signal structure of cooperative source, which is not present in previous studies. Throughout the oblation studies and real-world data based experiments, the effectiveness and efficiency are demonstrated.

CRediT authorship contribution statement

Hailun Chu: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - Original draft preparation. **Chao Li:** Data curation, Writing - Review & Editing, Investigation. **Haibin Wang:** Conceptualization, Investigation, Writing - Review & Editing, Supervision, Project administration, Funding acquisition. **Jun Wang:** Writing - Review & Editing, Resources. **Yupeng Tai:** Writing - Review & Editing, Resources. **Yonglin Zhang:** Writing - Review & Editing, Resources. **Lei Zhou:** Writing - Review & Editing, Software. **Fan Yang:** Writing - Review & Editing, Resources. **Yannick Benezeth:** Writing - Review & Editing, Resources, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence.

Data availability

The authors do not have permission to share data.

Acknowledgement

This research was supported in part by the China Scholarship Council under Grant No. 202110580001, National Natural Science Foundation of China under Grant No. 62171440 and No. 62301551, and CAS Specific Research Assistant Funding Program.

References

Arveson, P.T., Vendittis, D.J., 2000. Radiated noise characteristics of a modern cargo ship. The Journal of the Acoustical Society of America 107, 118–129.

Capon, J., 1969. High-resolution frequency-wavenumber spectrum analysis. Proceedings of the IEEE 57, 1408–1418.

Carey, W., Reese, J., Stuart, C., 1997. Mid-frequency measurements of array signal and noise characteristics. IEEE Journal of Oceanic Engineering 22, 548–565.

Chen, J., Mao, Q., Liu, D., 2020. Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation, in: Proc. Interspeech 2020, pp. 2642–2646.

Chi, C., Pallayil, V., Chitre, M., 2020. Design of an adaptive noise canceller for improving performance of an autonomous underwater vehicle-towed linear array. Ocean Engineering 202, 106886.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Chu, H., Li, C., Wang, H., Wang, J., Tai, Y., Zhang, Y., Yang, F., Benezeth, Y., 2023. A deep-learning based high-gain method for underwater acoustic signal detection in intensity fluctuation environments. Applied Acoustics 211, 109513.

Feng, J., Zou, N., Wang, Y., Hao, Y., 2018. Methods of suppressing tow ship noise with a horizontal linear array. The Journal of the Acoustical Society of America 143, 1959–1959.

Gershman, A., Turchin, V., Zverev, V., 1995. Experimental results of localization of moving underwater signal by adaptive beamforming. IEEE Transactions on Signal Processing 43, 2249–2257.

Godara, L.C., 1991. Adaptive postbeamformer interference canceler with improved performance in the presence of broadband directional sources. The Journal of the Acoustical Society of America 89, 266–273.

Han, Y., Li, Y., Liu, Q., Ma, Y., 2020. Deeplofargram: A deep learning based fluctuating dim frequency line detection and recovery. The Journal of the Acoustical Society of America 148, 2182–2194.

Hao, X., Su, X., Horaud, R., Li, X., 2021. Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement, in: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6633–6637.

Hua, W., Dai, Z., Liu, H., Le, Q., 2022. Transformer quality in linear time, in: International Conference on Machine Learning, PMLR. pp. 9099–9117.

Hui, J., Song, M., Li, J., 2018. Research on suppression for tow-ship interference. The Journal of the Acoustical Society of America 144, 1944–1944.

Ijsselmuide, S., Beerens, P., 2001. Adaptive beamforming algorithms for passive sonar arrays.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Lam, M.W.Y., Wang, J., Su, D., Yu, D., 2021. Effective low-cost time-domain audio separation using globally attentive locally recurrent networks, in: 2021 IEEE Spoken Language Technology Workshop (SLT), pp. 801–808.

Liang, J., Zhang, T., Xu, W., Zhao, H., 2023. A linear near-field interference cancellation method based on deconvolved conventional beamformer using fresnel approximation. IEEE Journal of Oceanic Engineering 48, 365–371.

Liu, J., Shi, D., Wu, G., 2015. Hybrid chaos optimization algorithm based on kent mapping. Computer Engineering and Design 6, 1498–1503.

Luo, Y., Chen, Z., Yoshioka, T., 2020. Dual-path rnn: Efficient long

sequence modeling for time-domain single-channel speech separation, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 46–50.

Luo, Y., Mesgarani, N., 2019. Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing 27, 1256–1266.

Lv, S., Hu, Y., Zhang, S., Xie, L., 2021. DCCRN+: Channel-Wise Subband DCCRN with SNR Estimation for Speech Enhancement, in: Proc. Interspeech 2021, pp. 2816–2820.

Mitra, S.K., 2001. Digital signal processing: a computer-based approach. McGraw-Hill Higher Education.

Narang, S., Chung, H.W., Tay, Y., Fedus, W., Fevry, T., Matena, M., Malkan, K., Fiedel, N., Shazeer, N., Lan, Z., et al., 2021. Do transformer modifications transfer across implementations and applications? arXiv preprint arXiv:2102.11972 .

Pandey, A., Wang, D., 2019. Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain, in: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6875–6879.

Peng, B., Alcaide, E., Anthony, Q., et al., 2023. Rwkv: Reinventing rnns for the transformer era. arXiv preprint arXiv:2305.13048 .

Rugini, L., Banelli, P., 2016. On the equivalence of maximum snr and mmse estimation: Applications to additive non-gaussian channels and quantized observations. IEEE Transactions on Signal Processing 64, 6190–6199.

Shazeer, N., 2020. Glu variants improve transformer. arXiv preprint arXiv:2002.05202 .

Shi, S.G., Li, Y., Zhu, Z.r., Shi, J., 2019. Real-valued robust doa estimation method for uniform circular acoustic vector sensor arrays based on worst-case performance optimization. Applied Acoustics 148, 495–502.

Shu, X., Wang, H., Yang, X., Wang, J., 2016. Chaotic modulations and performance analysis for digital underwater acoustic communications. Applied Acoustics 105, 200–208.

Smith, T.A., Rigby, J., 2022. Underwater radiated noise from marine vessels: A review of noise reduction methods and technology. Ocean Engineering 266, 112863.

Song, Y., Liur, F., Shen, T., 2022. Method of underwater acoustic signal denoising based on dual-path transformer network. IEEE Access , 1–1.

Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., Zhong, J., 2021. Attention is all you need in speech separation, in: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 21–25.

Turin, G., 1960. An introduction to matched filters. IRE Transactions on Information Theory 6, 311–329.

Van Veen, B., Buckley, K., 1988. Beamforming: a versatile approach to spatial filtering. IEEE ASSP Magazine 5, 4–24.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017a. Attention is all you need. Advances in neural information processing systems 30.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I., 2017b. Attention is all you need, in: Advances in Neural Information Processing Systems, Curran Associates, Inc.

Wilson, J.H., Nuttall, A.H., Prater, R.A., 2006. Noise suppression using the coherent onion peeler. The Journal of the Acoustical Society of America 120, 3627–3634.

Yan, S., Ma, Y., 2005. Robust supergain beamforming for circular array via second-order cone programming. Applied Acoustics 66, 1018–1032.

Yang, L., McKay, M.R., Couillet, R., 2018. High-dimensional mvdr beamforming: Optimized solutions based on spiked random matrix models. IEEE Transactions on Signal Processing 66, 1933–1947.

Zheng, C., Peng, X., Zhang, Y., Srinivasan, S., Lu, Y., 2021. Interactive speech and noise modeling for speech enhancement. Proceedings of the AAAI Conference on Artificial Intelligence 35, 14549–14557.

Zhuoran, S., Mingyuan, Z., Haiyu, Z., Shuai, Y., Hongsheng, L., 2021. Efficient attention: Attention with linear complexities, in: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 3530–3538.