Information Fusion 122 (2025) 103218

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

Cross-attention among spectrum, waveform and SSL representations with bidirectional knowledge distillation for speech enhancement☆

Hang Chen [a], Chenxi Wang [a], Qing Wang [a], Jun Du [a,*], Sabato Marco Siniscalchi [b], Genshun Wan [c], Jia Pan [c], Huijun Ding [d]

[a] University of Science and Technology of China, Hefei, 230026, Anhui, PR China
[b] University of Palermo, Palermo, 90133, Italy
[c] iFLYTEK Research, Hefei, 230088, Anhui, PR China
[d] Shenzhen University, Shenzhen, 518060, Guangdong, PR China

ARTICLE INFO

Keywords:
Speech enhancement
Cross-domain feature
Self-supervised learning
Knowledge distillation
Homoscedastic uncertainty

ABSTRACT

We have developed an innovative speech enhancement (SE) model backbone that utilizes cross-attention among spectrum, waveform and self-supervised learned representations (CA-SW-SSL) to integrate knowledge from diverse feature domains. The CA-SW-SSL model integrates the cross spectrum and waveform attention (CSWA) model to connect the spectrum and waveform branches, along with a dual-path cross-attention module to select outputs from different layers of the self-supervised learning (SSL) model. To handle the increased complexity of SSL integration, we introduce a bidirectional knowledge distillation (BiKD) framework for model compression. The proposed adaptive layered distance measure (ALDM) maximizes the Gaussian likelihood between clean and enhanced multi-level SSL features during the backward knowledge distillation (BKD) process. Meanwhile, in the forward process, the CA-SW-SSL model acts as a teacher, using the novel teacher–student Barlow Twins (TSBT) loss to guide the training of the CSWA student models, including both lite and tiny versions. Experiments on the DNS-Challenge and Voicebank+Demand datasets demonstrate that the CSWA-Lite+BiKD model outperforms existing joint spectrum-waveform methods and surpasses the state-of-the-art on the DNS-Challenge non-blind test set with half the computational load. Further, the CA-SW-SSL+BiKD model outperforms all CSWA models and current SSL-based methods.

1. Introduction

Speech enhancement (SE) aims to extract clean speech from signals primarily degraded by noise to improve speech quality and intelligibility [1]. However, traditional SE algorithms [2–4] often struggle to handle unexpected non-stationary noise in real-world conditions. In recent years, data-driven SE approaches [5–7] using deep neural networks (DNNs) [8] have gained increased attention. These DNN-based SE approaches can be broadly classified into spectrum-domain and waveform-domain methods based on the input features. Specifically, spectrum-domain methods [5,9] benefit from the short-time Fourier transform (STFT) guided by human expert knowledge, offering harmonic information and low temporal resolution. However, they may converge to suboptimal solutions due to difficulties compensating between magnitude and phase [10]. On the other hand, waveform-domain methods [11,12] directly map between noisy and clean waveforms, avoiding optimization issues related to magnitude-phase compensation. Nevertheless, waveform-based solutions cannot directly utilize harmonic information, which is crucial for speech quality. As a result, joint spectrum-waveform methods have been proposed [13–15], which utilize a single model to process both spectrum and waveform features. This integration leverages the complementary strengths of both domains, leading to significant performance improvements.

In addition to traditional spectrum and waveform features, self-supervised learning (SSL) features have recently captured substantial attention due to their exceptional performance and robust generalization capabilities. SSL involves pretrained models on unlabeled data to extract task-agnostic representations, which are then utilized as inputs for subsequent model training in the target task. SSL has proven highly effective in various downstream tasks, including automatic speech recognition [16], speaker recognition [16], and spoken

☆ This work was supported by the National Natural Science Foundation of China under Grant 62171427.
* Corresponding author.
E-mail addresses: hangchen@ustc.edu.cn (H. Chen), cx_wang@mail.ustc.edu.cn (C. Wang), qingwang2@ustc.edu.cn (Q. Wang), jundu@ustc.edu.cn (J. Du), sabatomarco.siniscalchi@unipa.it (S.M. Siniscalchi), gswan@iflytek.com (G. Wan), jiapan@iflytek.com (J. Pan), hjding@szu.edu.cn (H. Ding).

https://doi.org/10.1016/j.inffus.2025.103218
Received 25 September 2024; Received in revised form 26 March 2025; Accepted 10 April 2025
Available online 24 April 2025

language understanding [17]. A few studies have also explored integrating SSL features into SE by replacing [18] or supplementing [19] spectrum features, showing significant performance improvements. Recently, [20] has leveraged consistency-preserving loss and perceptual contrast stretching to boost SSL-based SE methods further.

Inspired by the achievements of the SSL-based and joint spectrum-waveform SE methods, this work proposes a novel joint spectrum-waveform-SSL model backbone for SE, fundamentally anchored in the cross-attention among spectrum, waveform and SSL representations. We refer to our proposed solution as CA-SW-SSL. Specifically, the CA-SW-SSL model evolves from the cross-spectrum and waveform attention (CSWA) model and introduces a dual-path cross-attention (DPCA) module for handling multi-layer SSL features. The CSWA model combines the classic spectrum-domain S-TCN [21] and waveform-domain ConvTasnet [12] models using two cross-attention modules with opposing directionals, facilitating efficient information exchange. DPCA effectively achieves dynamic alignment and layer selection by separately calculating time-wise and layer-wise attention across the outputs of all layers in the SSL model and then combining these attention weights using the Kronecker product. Experimental results prove the proposed CA-SW-SSL model surpasses existing SSL-based SE and joint spectrum-waveform methods. However, these remarkable performance gains come with a substantial demand for parameters and computational resources, underscoring the necessity to focus on model compression [22].

Knowledge distillation (KD) [23] is a widely used technique for model compression, where a smaller student model is trained under the supervision of a larger teacher model [24,25]. Several studies [26–28] have successfully applied KD to SE, achieving significant reductions in model parameters and computational complexity without noticeable performance degradation. Therefore, this work also proposes a novel bidirectional knowledge distillation (BiKD) framework, which is partitioned into backward and forward processes based on the direction of knowledge transfer. In the process of backward knowledge distillation (BKD), a novel adaptive layered distance measure (ALDM) is proposed to optimize the whole model by improving the Gaussian likelihood between the clean and enhanced multi-level features mapped by a frozen SSL model, with homoscedastic uncertainty [29] interpreted as layer-dependent weighting. For forward knowledge distillation (FKD), the CA-SW-SSL model takes on the role of the teacher, supervising the training of the CSWA model and its smaller versions, acting as student models. To achieve this, an innovative teacher–student Barlow Twins [30] (TSBT) loss is used to ensure that the cross-correlation matrix between the teacher and student intermediate features closely approximates the identity matrix. Finally, we obtained a highly efficient CSWA-Tiny model, with only $1/39$th of the parameters and $2/5$th of the computational complexity of the CA-SW-SSL model.

Key contributions of our work can be outlined as follows:

(1) Proposing a novel SE backbone featuring cross-attention among spectrum, waveform, and SSL representations (CA-SW-SSL), which includes cross spectrum and waveform attention (CSWA) for joint modeling and dual-path cross-attention (DPCA) for multi-layer SSL features. To the best of our knowledge, the CA-SW-SSL model is the first to combine spectrum, waveform, and SSL features in a single SE model.
(2) Developing a bidirectional knowledge distillation (BiKD) framework for model compression. An innovative adaptive layered distance measure (ALDM) performs the backward process by maximizing the Gaussian likelihood between clean and enhanced multi-level SSL features. Meanwhile, the forward process employs the CA-SW-SSL model as the teacher, utilizing the innovative teacher–student Barlow Twins (TSBT) loss to guide the training of the CSWA student models.
(3) Assessing the effectiveness and generalizability of CA-SW-SSL and BiKD through a set of comprehensive experiments. The CA-SW-SSL model achieves state-of-the-art (SOTA) results on the widely recognized DNS-Challenge and Voicebank+Demand datasets. Through the BiKD framework, reducing the parameters by $38/39$ and the computational complexity by $3/5$, the resulting CSWA-Lite model still demonstrates outstanding performance on both datasets.

The remainder of the paper is organized as follows. Section 2 introduces related works. In Section 3, the proposed CSWA and CA-SW-SSL models are illustrated. Section 4 describes our proposed BiKD framework. Section 5 discusses the experimental results and analysis. Finally, our findings are summarized in Section 6.

2. Related work

2.1. Joint spectrum-waveform SE

Spectrum-domain methods, e.g., [5,9,31] harness short-time Fourier transform (STFT), effectively leveraging crucial harmonic information for speech quality while maintaining reduced time resolution. Efficient context modeling is allowed by those models with minimal computational cost. However, these methods are prone to converging to local suboptimal solutions due to the compensation problem between magnitude and phase [10]. In contrast, waveform-domain methods directly model the waveform mapping [12,32], bypassing the compensation problem but sacrificing harmonic information.

Some researchers have proposed joint spectrum-waveform methods that integrate the strengths of both [13–15,33,34]. Specifically, TFT-Net [14] directly transforms the noisy complex spectrum into the corresponding clean waveform, effectively harnessing the harmonic information while circumventing the optimization challenges arising from the interplay between magnitude and phase. Similarly, MDPhD [13] integrates both spectrum-domain and waveform-domain models in a cascaded manner, introducing an auxiliary loss at the intermediate stage to harmonize the contributions of the two models. However, in both TFT-Net and MDPhD, the spectrum and waveform branches are arranged serially, constraining the exchange of information between different representation domains. To address this, DBNet [15] introduces a novel dual-branch structure with alternating interconnections, where a novel bridge layer is employed to facilitate information exchange between spectrum and waveform branches. A similar dual-branch structure is found in WSFNet [34], where dual-path RNN blocks are embedded into the bottleneck layer to model intra-frame and inter-frame long-range contextual correlations. Further, WMPNet [33] builds upon the spectrum-waveform dual-branch structure by incorporating a fusion sub-network to merge the two pre-enhanced speech.

It is worth noting that the above techniques all use the same architecture for processing both the spectral- and waveform-based inputs, disregarding their unique properties. Additionally, the information exchange between the spectrum and waveform branches is often one-way, significantly limiting the overall performance.

2.2. SSL-based SE

The SSL-based SE method, which traces its origins back to [18], involves using the final output of an SSL model as input to a three-layer bidirectional long short-term memory (BiLSTM) network for estimating the ideal non-negative phase sensitive mask [35]. However, replacing the spectrum feature with SSL features did not yield significant benefits. Therefore, [19] proposed using SSL features as auxiliary inputs, combining them with the $\log 1p$ [36] feature to predict the magnitude mask by a two-layer BiLSTM, demonstrating superior performance compared to the replacement method. In [20], consistency-preserving loss and contrast stretching were introduced to boost the overall speech quality further.

Inspired by deep feature loss [37], several studies have proposed loss functions in the feature spaces of the SSL models. One such example is the phone-fortified perceptual loss (PFPL) [38], defined as the Wasserstein distance [39] on latent representations of the wav2vec model [40] thereby incorporating phonetic information for training SE models. [41] introduced the mean squared error (MSE) between the enhanced and clean intermediate encoded features from SSL models as the optimization target, showing a strong correlation with speech quality and intelligibility. More recently, [42] proposed a contrastive regularization technique, which utilizes contrastive learning to minimize the distance between clean and enhanced speech while maximizing the distance between noisy and enhanced speech within the representation space of the SSL model.

Despite the above encouraging efforts, additional investigation is needed to understand the specific information that SE models gain from SSL models to enhance performance, as well as to identify which SSL model layer provides the optimal auxiliary features for forward inference and the best representation space for loss computation.

2.3. Knowledge distillation in SE

Knowledge distillation [23] (KD) refers to the process where a smaller student model is trained under the guidance of a larger teacher model [24,25]. The challenges of defining the specific form of knowledge and selecting the appropriate distillation loss directly impact the effectiveness of knowledge transfer from teacher to student. While KD was initially applied at the output layer [43,44] of deep models, recent works have honed in on distillation at intermediate layers [45–47]. Moreover, it is widely acknowledged that two of the most effective distillation loss functions are MSE [45,46,48,49], and mean absolute error (MAE) [43].

In SE, there exist a few studies leveraging KD. For example, MV-AT [26] applied attention transfer [46] in waveform-domain SE. In [27], a cross-layer connection strategy that combines multi-level information from the teacher and transfers it to the student using a frame-level similarity distillation loss was explored. ABC-KD [28] combines KD with a layer-wise cross-attention mechanism to compress knowledge from multiple teacher layers into a single student layer. The two-step KD method [50] first pre-trains the student using a fine-grained KD loss to match the student's intra-activation Gram matrices to the teacher, and then it transits to a supervised training regime.

3. Cross-attention among Spectrum, Waveform and SSL Representations

This section presents the CA-SW-SSL model, which enhances the CSWA model by incorporating a dual-path cross-attention, DPCA, module to introduce multi-layer SSL features effectively.

3.1. Cross spectrum and waveform attention

Fig. 1 shows the Cross Spectrum and Waveform Attention, CSWA, model, which is composed of five distinct components: Spectrum branch (in purplish), waveform branch (in pink), context module (in beige), and two cross-attention modules with opposing directions (in gray). Given a noisy waveform $\boldsymbol{x} \in \mathbb{R}^L$, the encoder in the waveform branch, which consists of N_{W} stacked 1D convolutional blocks (Conv1Ds), outputs the 2D representation $Z_i \in \mathbb{R}^{T_{\mathrm{W}} \times C_{\mathrm{W}}}$. The waveform encoding process can be described as follows:

$$Z = \mathrm{Conv1D}_1\left(\cdots \mathrm{Conv1D}_{N_{\mathrm{W}}}(\boldsymbol{x})\right) \tag{1}$$

where L is the length of the waveform. T_{W} and C_{W} represent the sequence length and the channel number of the encoded waveform embedding, respectively. The Conv1D is similar to the basic unit in ConvTasNet [12] and includes a 1×1 convolution (1×1-Conv) followed by a depthwise convolution (D-Conv), which projects the input to a predetermined C_{W}-channel space for decoupling noise and speech. Additionally, two 1×1-Convs serve as the residual path and the skip-connection path, respectively. For clarity, some normalization layers and activation functions have been omitted in our description, but details can be found in [12].

The spectrum encoder, which consists of N_{S} stacked light-weight temporal convolution modules (TCMs) [21], learns the deep representation $Y \in \mathbb{R}^{T_{\mathrm{S}} \times C_{\mathrm{S}}}$ from log-power spectra (LPS) features. The spectrum encoding process is described as follows:

$$Y = \mathrm{TCM}_1\left(\cdots \mathrm{TCM}_{N_{\mathrm{S}}}\left(\log |X|^2\right)\right) \tag{2}$$

where $X \in \mathbb{C}^{T_{\mathrm{S}} \times C_0}$ is the noisy spectrum, which is obtained by applying the STFT to the noisy waveform $\boldsymbol{x}$, with T_{S} and C_0 denoting the number of frames and frequency bins in the spectrogram, respectively. C_{S} is the number of channels in the spectrum encoder. Every TCM comprises two 1×1-Convs and two dilated depthwise convolutions (DD-Convs) with a gating mechanism. The first 1×1-Conv compresses the input into a more compact channel space, specifically to $C_{\mathrm{S}}/4$ channels. Subsequently, the gated DD-Convs are applied, where a regular DD-Conv is multiplied by another DD-Conv, with the sigmoid function scaling the output values into the range $(0, 1)$. Finally, the output 1×1-Conv restores the dimensionality to C_{S} channels.

The spectrum embedding Y offers low temporal resolution and integrates expert knowledge but lacks phase information. In contrast, the waveform embedding Z contains both magnitude and phase information but lacks useful harmonic content. To leverage the complementary nature of these embeddings, we introduce a waveform-to-spectrum (W2S) cross-attention module, where Y serves as the query input, Z serves as the key and value inputs, for integrating phase information from the waveform branch into the spectrum branch. The detailed process is as follows:

$$A_n^{\mathrm{W2S}} = \mathrm{SoftMax}\left(\frac{YQ_n^{\mathrm{W2S}}\left(ZK_n^{\mathrm{W2S}}\right)^{\top}}{\sqrt{C_{\mathrm{W2S}}}}\right) \tag{3}$$

$$D_n^{\mathrm{W2S}} = A_n^{\mathrm{W2S}} Z V_n^{\mathrm{W2S}} \tag{4}$$

$$E = \mathrm{Concat}\left(D_1^{\mathrm{W2S}}, \ldots, D_{N_{\mathrm{W2S}}}^{\mathrm{W2S}}, Y\right) O^{\mathrm{W2S}} \tag{5}$$

where $E \in \mathbb{R}^{T_{\mathrm{S}} \times C_{\mathrm{S}}}$ and $O^{\mathrm{W2S}} \in \mathbb{R}^{(N_{\mathrm{W2S}} C_{\mathrm{W2S}} + C_{\mathrm{S}}) \times C_{\mathrm{S}}}$ denote the fused spectrum embedding and the projection matrices of the output, respectively. N_{W2S} attention heads are used, and $n \in \{1, 2, \ldots, N_{\mathrm{W2S}}\}$ is the index of the head. For the nth attention head, $D_n^{\mathrm{W2S}} \in \mathbb{R}^{T_{\mathrm{S}} \times C_{\mathrm{W2S}}}$ and $A_n^{\mathrm{W2S}} \in \mathbb{R}^{T_{\mathrm{S}} \times T_{\mathrm{W}}}$ denote the output and the attention weight, respectively. $Q_n^{\mathrm{W2S}} \in \mathbb{R}^{C_{\mathrm{S}} \times C_{\mathrm{W2S}}}$, $K_n^{\mathrm{W2S}} \in \mathbb{R}^{C_{\mathrm{W}} \times C_{\mathrm{W2S}}}$ and $V_n^{\mathrm{W2S}} \in \mathbb{R}^{C_{\mathrm{W}} \times C_{\mathrm{W2S}}}$ are the projection matrices of query, key and value. C_{W2S} is the number of channel in the W2S cross-attention module.

Next, the fused spectrum embedding E first undergoes local context modeling through the N_{L} stacked TCMs, generating the local embedding $H \in \mathbb{R}^{T_{\mathrm{S}} \times C_{\mathrm{L}}}$ to capture short-term dependencies. Subsequently, an N_{G}-layer Conformer [51] takes over, modeling global contexts and incorporating long-term dependencies to obtain the global embedding $U \in \mathbb{R}^{T_{\mathrm{S}} \times C_{\mathrm{G}}}$. The context modeling process is:

$$H = \mathrm{TCM}_1\left(\cdots \mathrm{TCM}_{N_{\mathrm{L}}}(E)\right) \tag{6}$$

$$U = \mathrm{Conformer}_1\left(\cdots \mathrm{Conformer}_{N_{\mathrm{G}}}(H)\right) \tag{7}$$

where C_{L} and C_{G} are the channel settings of the local-contextual and global-contextual modules, respectively. Each Conformer is composed of four modules stacked together, i.e, a feed-forward module, a self-attention module, a convolution module, and a second feed-forward module in the end. More details can be found in [51].

Finally, the global embedding U is fed to the spectrum decoder, which mirrors the spectrum encoder and thereby consists of N_{S} TCMs for predicting complex mask $\widehat{M} \in \mathbb{C}^{T_{\mathrm{S}} \times C_0}$. Notably, considering the differences between the real and imaginary parts, the entire decoder is

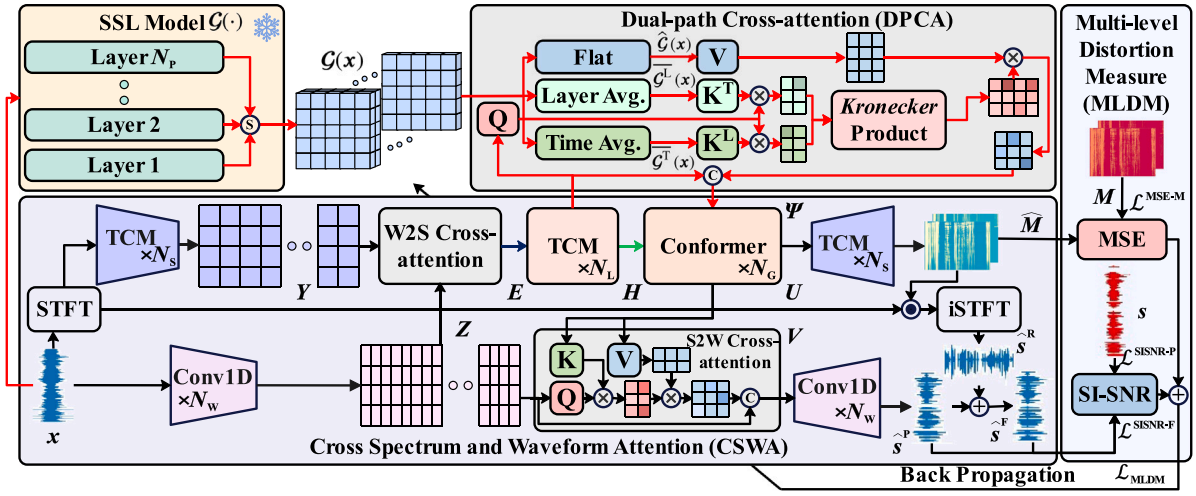

Fig. 1. Illustration of the proposed CA-SW-SSL model, which adapts the CSWA model for cross-domain information exchange between the spectrum and waveform branches and employs a DPCA module to select SSL features from different layers. Finally, the MLDM loss supervises the individual branch and the fused outputs. $\boldsymbol{x}$, $\boldsymbol{s}$, $\hat{\boldsymbol{s}}^{\mathrm{P}}$, $\hat{\boldsymbol{s}}^{\mathrm{R}}$, and $\hat{\boldsymbol{s}}^{\mathrm{F}}$ denote noisy, clean, waveform-branch predicted, the spectrum-branch reconstructed, and the fused waveforms, respectively. The cIRM is denoted by M. Y, Z, E, H, Ψ, U, and V denote the noisy spectral, noisy waveform, fused, local context, second-fused, global context, target spectral, and target waveform embeddings, respectively. $\mathcal{G}(\boldsymbol{x})$ denotes the stacked outputs from all layers of the SSL model $\mathcal{G}(\cdot)$. $\hat{\mathcal{G}}(\boldsymbol{x})$, $\overline{\mathcal{G}^{\mathrm{L}}}(\boldsymbol{x})$, and $\overline{\mathcal{G}^{\mathrm{T}}}(\boldsymbol{x})$ represent the flattened, layer-averaged, and time-averaged versions of $\mathcal{G}(\boldsymbol{x})$, respectively.

designed as two parallel mirrored sections to predict them separately. The detailed process is as follows:

$$\begin{aligned}\widehat{M} =& \mathrm{TCM}_1(\cdots \mathrm{TCM}_{N_S/2}(U)) \\ &+ i\mathrm{TCM}_1(\cdots \mathrm{TCM}_{N_S/2}(U))\end{aligned} \tag{8}$$

Then, the predicted mask $\widehat{M}$ is used to filter the noisy spectrum X and reconstruct the waveform $\hat{s} \in \mathbb{R}^L$ by inverse short-time Fourier transform (iSTFT). The reconstruction process is briefly described as follows:

$$\hat{\boldsymbol{s}}^{\mathrm{R}} = \mathrm{iSTFT}(\widehat{M} \cdot X) \tag{9}$$

The global embedding U is fed back to the waveform branch by a spectrum-to-waveform (S2 W) cross-attention module. Specifically, the S2 W cross-attention module includes N_{S2W} attention heads and takes the waveform embedding Z as the query input, with the global representation U serving as the key and value inputs. The detailed feedback process is as follows:

$$A_n^{\mathrm{S2W}} = \mathrm{SoftMax}\left(\frac{ZQ_n^{\mathrm{S2W}}(UK_n^{\mathrm{S2W}})^{\mathrm{T}}}{\sqrt{C_{\mathrm{S2W}}}}\right) \tag{10}$$

$$D_n^{\mathrm{S2W}} = A_n^{\mathrm{S2W}} U V_n^{\mathrm{S2W}} \tag{11}$$

$$V = \mathrm{Concat}\left(D_1^{\mathrm{S2W}}, \ldots, D_{N_{\mathrm{S2W}}}^{\mathrm{S2W}}, Z\right) O^{\mathrm{S2W}} \tag{12}$$

where $V \in \mathbb{R}^{T_{\mathrm{W}} \times C_{\mathrm{W}}}$ and $O^{\mathrm{S2W}} \in \mathbb{R}^{(N_{\mathrm{S2W}} C_{\mathrm{S2W}} + C_{\mathrm{W}}) \times C_{\mathrm{W}}}$ represent the fused waveform embedding and the output projection matrices, respectively. For the nth attention head, $D_n^{\mathrm{S2W}} \in \mathbb{R}^{T_{\mathrm{W}} \times C_{\mathrm{S2W}}}$ and $A_{i,n}^{\mathrm{S2W}} \in \mathbb{R}^{T_{\mathrm{W}} \times T_{\mathrm{S}}}$ denote the output and the attention weight, respectively. $Q_n^{\mathrm{S2W}} \in \mathbb{R}^{C_{\mathrm{W}} \times C_{\mathrm{S2W}}}$, $K_n^{\mathrm{S2W}} \in \mathbb{R}^{C_{\mathrm{G}} \times C_{\mathrm{S2W}}}$ and $V_n^{\mathrm{S2W}} \in \mathbb{R}^{C_{\mathrm{G}} \times C_{\mathrm{S2W}}}$ are the projection matrices of query, key and value, respectively. C_{S2W} is the number of channels in the S2 W cross-attention module. The S2 W cross-attention module transfers long-term context information and harmonic details from the STFT to the waveform branch, serving as a complementary input to the waveform embedding.

The waveform decoder mirrors the waveform encoder and thereby consists of N_{W} stacked Conv1Ds. It predicts the enhanced waveform $\hat{\boldsymbol{s}}^{\mathrm{P}} \in \mathbb{R}^L$ from the fused waveform embedding V using a 1D transposed convolution (Conv1DT):

$$\hat{\boldsymbol{s}}^{\mathrm{P}} = \mathrm{Conv1DT}\left(\mathrm{Conv1D}_1\left(\cdots \mathrm{Conv1D}_{N_{\mathrm{W}}}(V)\right)\right) \tag{13}$$

Finally, the waveform $\hat{\boldsymbol{s}}^F \in \mathbb{R}^L$ is generated by an equal combination of the output from the spectrum and waveform branches:

$$\hat{\boldsymbol{s}}^F = \hat{\boldsymbol{s}}^{\mathrm{R}} + \hat{\boldsymbol{s}}^{\mathrm{P}} \tag{14}$$

3.2. Multi-level distortion measure

Regarding the multiple outputs from the CSWA model, we use a multi-level distortion measure (MLDM) [52] as the overall loss function, which provides comprehensive and aligned supervision by designing specific loss components for each output to measure distortions accurately. Specifically, the scale-invariant signal-to-noise ratio (SISNR) is adopted to measure the distortions of the predicted waveforms $\hat{\boldsymbol{s}}^P$ and the fused waveforms $\hat{\boldsymbol{s}}^F$ compared to the clean waveforms $\boldsymbol{s} \in \mathbb{R}^L$. These loss components, referred to as SISNR-P and SISNR-F, are computed using the following formula:

$$\tilde{s}_\tau^{\mathrm{P/F}} = \hat{s}_\tau \left(\sum_{\tau=0}^{L-1} s_\tau^2\right) / \left(\sum_{\tau=0}^{L-1} \hat{s}_\tau^{\mathrm{P/F}} s_\tau\right) \tag{15}$$

$$\overline{\mathcal{L}}^{\mathrm{SISNR\text{-}P/F}} = -10\log \frac{\sum_{\tau=1}^{L} s_\tau^2}{\sum_{\tau=1}^{L} \left(\tilde{s}_\tau^{\mathrm{P/F}} - s_\tau\right)^2} \tag{16}$$

where $\hat{s}_\tau^{\mathrm{P}}$, $\hat{s}_\tau^{\mathrm{F}}$ and s_τ are values at the τ-th time step of the predict waveform $\hat{\boldsymbol{s}}^{\mathrm{P}}$, the fused waveform $\hat{\boldsymbol{s}}^{\mathrm{F}}$ and the clean waveform $\boldsymbol{s}$, respectively. At the same time, the MSE between the complex ideal ratio mask (cIRM) [53] $M \in \mathbb{C}^{T_{\mathrm{S}} \times C_0}$ and the predicted complex mask $\hat{M}_i$ is used as the spectral similarity loss. These loss components, denoted as MSE-M, can be computed as follows:

$$\overline{\mathcal{L}}^{\mathrm{MSE\text{-}M}} = \frac{1}{T_{\mathrm{S}} C_0} \sum_{t=1}^{T_{\mathrm{S}}} \sum_{j=1}^{C_0} \left|\hat{m}_{t,j} - m_{t,j}\right|^2 \tag{17}$$

where $\hat{m}_{t,j}$ and $m_{t,j}$ are the complex values at the tth frame and jth frequency bin of the predict mask $\hat{M}$ and the cIRM M, respectively. However, the raw values of MSE-M and SISNR-P exhibit a different order of magnitude; therefore, a scaling procedure is employed:

$$\widetilde{\mathcal{L}}^{\mathrm{MSE\text{-}M}} = \delta \overline{\mathcal{L}}^{\mathrm{MSE\text{-}M}} = 10^{\left\lfloor \log_{10} \frac{|\overline{\mathcal{L}}^{\mathrm{SISNR\text{-}P}}|}{|\overline{\mathcal{L}}^{\mathrm{MSE\text{-}M}}|} \right\rfloor} \overline{\mathcal{L}}^{\mathrm{MSE\text{-}M}} \tag{18}$$

where δ is the zoom factor and $\lfloor \cdot \rfloor$ denotes the floor function. When computing the gradient, the zoom factor δ is treated as a constant,

ensuring that the magnitudes of $\widetilde{\mathcal{L}}^{\text{MSE-M}}$ and $\widetilde{\mathcal{L}}^{\text{SISNR-P}}$ are comparable. Finally, the MLDM $\mathcal{L}^{\text{MLDM}}$, defined as a weighted combination of MSE-M, SISNR-P, and SISNR-F, is calculated as follows:

$$\mathcal{L}^{\text{MLDM}} = \widetilde{\mathcal{L}}^{\text{MSE-M}} + \overline{\mathcal{L}}^{\text{SISNR-P}} + \overline{\mathcal{L}}^{\text{SISNR-F}} \tag{19}$$

where the contribution of each branch and their fusion is considered equal in this study.

3.3. DPCA with multi-layer SSL features

Given an SSL model $\mathcal{G}(\cdot)$ with N_{P} layers, all candidate SSL features can be denoted as $\mathcal{G}(\boldsymbol{x}) \in \mathbb{R}^{N_{\text{P}} \times T_{\text{P}} \times C_{\text{P}}}$, where T_{P} and C_{P} represent the sequence length and the channel number, respectively. Compared with the W2S and S2 W cross-attention modules, the DPCA module offers a significant improvement by integrating both temporal- and layer-wise attention weights for selecting and fusing relevant SSL features across different layers. Specifically, DPCA first averages $\mathcal{G}(\boldsymbol{x})$ along the layer dimension to obtain $\overline{\mathcal{G}^{\text{L}}}(\boldsymbol{x}) \in \mathbb{R}^{T_{\text{P}} \times C_{\text{P}}}$. Then, the local context embedding H and $\overline{\mathcal{G}^{\text{L}}}$ serve as query and key inputs to generate the temporal-wise attention matrix $A_n^{\text{T}} \in \mathbb{R}^{T_{\text{S}} \times T_{\text{P}}}$ as follows:

$$A_n^{\text{T}} = \text{SoftMax}\left(\frac{HQ_n \left[\overline{\mathcal{G}^{\text{L}}}(\boldsymbol{x}) K_n^{\text{T}}\right]^{\top}}{\sqrt{C_{\text{DP}}}} \right) \tag{20}$$

where $Q_n \in \mathbb{R}^{C_{\text{L}} \times C_{\text{DP}}}$ is the query projection matrix. $K_n^{\text{T}} \in \mathbb{R}^{C_{\text{P}} \times C_{\text{DP}}}$ represents the projection matrix of key in the temporal path. C_{DP} is the number of channels in the DPCA module.

In terms of layer-wise attention, DPCA first averages $\mathcal{G}(\boldsymbol{x})$ along the temporal axis to produce $\overline{\mathcal{G}^{\text{T}}}(\boldsymbol{x}) \in \mathbb{R}^{N_{\text{P}} \times C_{\text{P}}}$ as the query input, while retaining H as the key input. The resulting layer-wise attention matrix $A_n^{\text{L}} \in \mathbb{R}^{T_{\text{S}} \times N_{\text{P}}}$ is calculated as follows:

$$A_n^{\text{L}} = \text{SoftMax}\left(\frac{HQ_n \left[\overline{\mathcal{G}^{\text{T}}}(\boldsymbol{x}) K_n^{\text{L}}\right]^{\top}}{\sqrt{C_{\text{DP}}}} \right) \tag{21}$$

where $K_n^{\text{L}} \in \mathbb{R}^{C_{\text{P}} \times C_{\text{DP}}}$ represent the projection matrices of key in the layer path.

Then, the Kronecker product is adopted to combine the temporal- and layer-wise attention matrices, yielding a dual-path attention matrix $A_n^{\text{DP}} \in \mathbb{R}^{T_{\text{S}} \times N_{\text{P}} T_{\text{P}}}$ as follows:

$$A_n^{\text{DP}} = \left[\left(\boldsymbol{a}_{n,1}^{\text{L}} \otimes \boldsymbol{a}_{n,1}^{\text{T}} \right)^{\top}, \ldots, \left(\boldsymbol{a}_{n,T_{\text{S}}}^{\text{L}} \otimes \boldsymbol{a}_{n,T_{\text{S}}}^{\text{T}} \right)^{\top} \right]^{\top} \tag{22}$$

where $\boldsymbol{a}_{n,t}^{\text{L}} \in \mathbb{R}^{1 \times N_{\text{P}}}$ and $\boldsymbol{a}_{n,t}^{\text{T}} \in \mathbb{R}^{1 \times T_{\text{P}}}$ are the row vectors of A_n^{L} and A_n^{T}, respectively. Notably, the Kronecker product preserves the row-sum property, ensuring that each row still sums to 1.

Finally, A_n^{DP} is applied to $\widehat{\mathcal{G}}(\boldsymbol{x}) \in \mathbb{R}^{N_{\text{P}} T_{\text{P}} \times C_{\text{P}}}$, which is the flattened version of $\mathcal{G}(\boldsymbol{x})$ along the channel dimension, for output the second-fused embedding $\Psi \in \mathbb{R}^{T_{\text{S}} \times C_{\text{L}}}$:

$$D_n^{\text{DP}} = A_n^{\text{DP}} \widehat{\mathcal{G}}(\boldsymbol{x}) \widehat{V}_n^{\text{DP}} \tag{23}$$

$$\Psi = \text{Concat}\left(D_1^{\text{DP}}, \ldots, D_{N^{\text{DP}}}^{\text{DP}}, H \right) O^{\text{DP}} \tag{24}$$

where $O^{\text{DP}} \in \mathbb{R}^{(N_{\text{DP}} C_{\text{DP}} + C_{\text{L}}) \times C_{\text{L}}}$ represents the output projection matrix. $D_n^{\text{DP}} \in \mathbb{R}^{T_{\text{S}} \times C_{\text{DP}}}$ and $V_n^{\text{DP}} \in \mathbb{R}^{C_{\text{P}} \times C_{\text{DP}}}$ denote the output and the value projection matrix of the nth attention head, respectively. N^{DP} is the number of attention heads in the DPCA module. In the CA-SW-SSL model, the fused representation Ψ replaces H in the subsequent process. Notably, the parameters of $\mathcal{G}(\cdot)$ are not updated during training.

4. Bidirectional knowledge distillation

Integrating the SSL model in CA-SW-SSL results in a substantial increase in parameters and computational load, posing significant challenges for real-world applications of the CA-SW-SSL model. In response to this challenge, we present an innovative bidirectional knowledge distillation (BiKD) framework for streamlining the SE model's complexity while keeping the same SE quality. BiKD consists of a loss function with forward and backward processes, namely:

$$\mathcal{L}^{\text{BiKD}} = \mathcal{L}^{\text{BKD}} + \mathcal{L}^{\text{FKD}} \tag{25}$$

where $\mathcal{L}^{\text{BKD}}$ and $\mathcal{L}^{\text{FKD}}$ represent the backward and forward distillation losses, respectively (see Fig. 2).

4.1. Backward knowledge distillation

Backward knowledge distillation (BKD) aims to transfer knowledge from SSL during the backward propagation process. Specifically, the enhanced waveform $\hat{\boldsymbol{s}} \in \{\hat{\boldsymbol{s}}^{\text{P}}, \hat{\boldsymbol{s}}^{\text{R}}, \hat{\boldsymbol{s}}^{\text{F}}\}$ and the target waveform $\boldsymbol{s}$ are processed by the SSL model $\mathcal{G}(\cdot)$ to obtain the enhanced and clean SSL representations, $\mathcal{G}(\hat{\boldsymbol{s}})$ and $\mathcal{G}(\boldsymbol{s})$, respectively. A baseline method for gauging the disparity between the enhanced and target SSL representations is to utilize the MSE loss. The MSE-SSL loss can be straightforwardly computed as follows:

$$\begin{aligned} \overline{\mathcal{L}}^{\text{MSE-SSL}} &= \frac{1}{N_{\text{P}}} \sum_{k=1}^{N_{\text{P}}} \overline{\mathcal{L}}_k^{\text{MSE-SSL}} \\ &= \frac{1}{N_{\text{P}} T_{\text{B}} C_P} \sum_{k=1}^{N_{\text{P}}} \sum_{t=1}^{T_{\text{B}}} \sum_{j=1}^{C_P} \left[\mathcal{G}(\hat{\boldsymbol{s}})_{k,t,j} - \mathcal{G}(\boldsymbol{s})_{k,t,j} \right]^2 \end{aligned} \tag{26}$$

where $\overline{\mathcal{L}}_k^{\text{MSE-SSL}}$ is the MSE value calculated with the kth layer SSL representations. Accordingly, by replacing SISNR and MSE-M in MLDM with the MSE-SSL loss, the BKD loss can be formulated as:

$$\mathcal{L}^{\text{BKD-MSE}} = \mathcal{L}^{\text{MSE-SSL-P}} + \mathcal{L}^{\text{MSE-SSL-R}} + \mathcal{L}^{\text{MSE-SSL-F}} \tag{27}$$

where $\mathcal{L}^{\text{MSE-SSL-P}}$, $\mathcal{L}^{\text{MSE-SSL-R}}$ and $\mathcal{L}^{\text{MSE-SSL-F}}$ represent the MSE-SSL loss of predicted $\hat{\boldsymbol{s}}^{\text{P}}$, reconstructed $\hat{\boldsymbol{s}}^{\text{R}}$ and fused $\hat{\boldsymbol{s}}^{\text{F}}$ waveform, respectively.

MSE-SSL presupposes that the distances across each layer's SSL representation space carry equal significance. However, in practice, different layers may vary in their importance. To explore this possibility, we propose a novel adaptive layered distance measure (ALDM), defined as the negative log-likelihood between the enhanced and clean SSL representations, as follows:

$$\mathcal{L}^{\text{ALDM}} = -\mathbf{log} p\left[\mathcal{G}(\boldsymbol{s}) \,|\, \mathcal{G}(\hat{\boldsymbol{s}}) \right] \tag{28}$$

Next, by leveraging the conditional independence assumption, $p[\mathcal{G}(\boldsymbol{s})|\mathcal{G}(\hat{\boldsymbol{s}})]$ can be effectively factorized across the layers, results in the following multi-layer likelihood expression:

$$\begin{aligned} p\left[\mathcal{G}(\boldsymbol{s}) \,|\, \mathcal{G}(\hat{\boldsymbol{s}}) \right] &= p\left[\mathcal{G}(\boldsymbol{s})_1, \ldots, \mathcal{G}(\boldsymbol{s})_{N_{\text{P}}} \,|\, \mathcal{G}(\hat{\boldsymbol{s}}) \right] \\ &= \prod_{k=1}^{N_{\text{P}}} p\left[\mathcal{G}(\boldsymbol{s})_k \,|\, \mathcal{G}(\hat{\boldsymbol{s}})_k \right] \end{aligned} \tag{29}$$

where $\mathcal{G}(\boldsymbol{s})_k \in \mathbb{R}^{T_{\text{P}} \times C_{\text{P}}}$ and $\mathcal{G}(\hat{\boldsymbol{s}})_k \in \mathbb{R}^{T_{\text{P}} \times C_{\text{P}}}$ denote the k-layer components of enhanced and clean SSL representations, respectively. The k-layer likelihood $p[\mathcal{G}(\boldsymbol{s})_k|\mathcal{G}(\hat{\boldsymbol{s}})_k]$ is defined as a Gaussian with mean given by $\mathcal{G}(\hat{\boldsymbol{s}})_k$ and an trainable noise parameter σ_k:

$$\begin{aligned} p\left[\mathcal{G}(\boldsymbol{s})_k \,|\, \mathcal{G}(\hat{\boldsymbol{s}})_k \right] &\sim \mathcal{N}\left(\mathcal{G}(\hat{\boldsymbol{s}})_k, \sigma_k^2 \right) \\ &= \frac{1}{\sqrt{2\pi}\sigma_k} \mathbf{exp}\left[-\frac{\left\| \mathcal{G}(\boldsymbol{s})_k - \mathcal{G}(\hat{\boldsymbol{s}})_k \right\|^2}{2\sigma_k^2} \right] \end{aligned} \tag{30}$$

Using Eq. (29) and (30), Eq. (28) can be rewritten as:

$$\mathcal{L}^{\text{ALDM}} = -\sum_{k=1}^{N_{\text{P}}} \mathbf{log} p[\mathcal{G}(\boldsymbol{s})_k | \mathcal{G}(\hat{\boldsymbol{s}})_k]$$

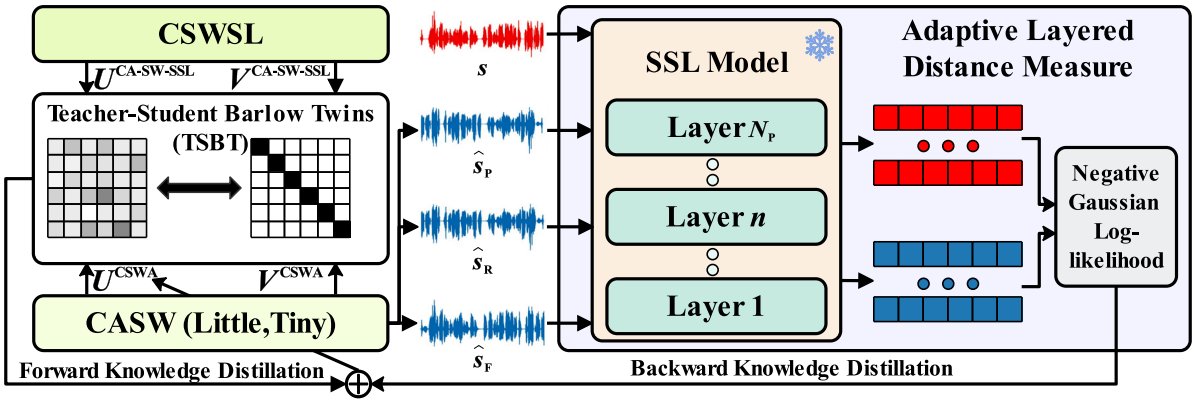

Fig. 2. Illustration of the proposed BiKD framework. The ALDM loss implements the backward process where s, $\hat{s}^{\text{P}}$, $\hat{s}^{\text{R}}$, and $\hat{s}^{\text{F}}$ denote clean, waveform-branch predicted, the spectrum-branch reconstructed, and the fused waveforms, respectively. The forward process adopts the TSBT loss, where U and V denote target spectral, and target waveform embeddings, respectively.

$$\propto \sum_{k=1}^{N_{\text{P}}} [\frac{1}{2\sigma_k^2} \|\mathcal{G}(s)_k - \mathcal{G}(\hat{s})_k\|^2 + \log\sigma_k]$$
$$= \sum_{k=1}^{N_{\text{P}}} \frac{1}{2\sigma_k^2} \overline{\mathcal{L}}_k^{\text{MSE-SSL}} + \log \prod_{k=1}^{N_{\text{P}}} \sigma_k \tag{31}$$

where σ_k can be interpreted as the homoscedastic uncertainty [29], serving as an adaptive layer-dependent weighting factor. As σ_k increases, the contribution of $\overline{\mathcal{L}}_k^{\text{MSE-SSL}}$ to the overall loss is proportionally reduced. Accordingly, the loss of BKD can be rewritten as follows:

$$\mathcal{L}^{\text{BKD-ALDM}} = \mathcal{L}^{\text{ALDM-R}} + \mathcal{L}^{\text{ALDM-P}} + \mathcal{L}^{\text{ALDM-F}} \tag{32}$$

where $\mathcal{L}^{\text{ALDM-P}}$, $\mathcal{L}^{\text{ALDM-R}}$ and $\mathcal{L}^{\text{ALDM-F}}$ represent the ALDM loss of the predicted waveform $\hat{s}^{\text{P}}$, the reconstructed waveform $\hat{s}^{\text{R}}$ and the fused waveform $\hat{s}^{\text{F}}$, respectively.

4.2. Forward knowledge distillation

Forward knowledge distillation (FKD) aims to transfer knowledge from the CA-SW-SSL model, as the teacher model, to the student models, including CSWA and its lite and tiny versions. Since the final output, which is traditionally used in KD, contains limited information, we propose to use the outputs of the global context and S2 W cross-attention modules, specifically U and V, as knowledge sources.

$\forall\Theta \in \{U, V\}$, the MSE between the corresponding internal outputs of the teacher and student models is employed to quantify the knowledge transfer:

$$\overline{\mathcal{L}}^{\text{MSE-TS}} = \frac{1}{T_{\text{B}} C_\Theta} \sum_{t=1}^{T_{\text{B}}} \sum_{j=1}^{C_\Theta} \left(\Theta_{t,j}^{\text{CA-SW-SSL}} - \Theta_{t,j}^{\text{CSWA}}\right)^2 \tag{33}$$

where C_Θ is the corresponding channel number. Therefore, the baseline FKD loss can be expressed as a sum of the MSE-TS losses from the global context module (U) and the S2 W cross-attention module (V):

$$\mathcal{L}^{\text{FKD}} = \overline{\mathcal{L}}^{\text{MSE-TS-U}} + \overline{\mathcal{L}}^{\text{MSE-TS-V}} \tag{34}$$

where $\mathcal{L}^{\text{MSE-TS-U}}$ and $\mathcal{L}^{\text{MSE-TS-V}}$ represent the MSE-TS loss of the global embedding U and the fused waveform embedding V, respectively.

The MSE loss emphasizes absolute numerical differences and overall smoothness between embeddings. Due to the significant difference in the number of network layers, it has been observed that the absolute numerical differences are challenging to reduce during training. Furthermore, the overall smoothness conceals local distortions, which can be amplified in subsequent network layers, ultimately deteriorating the final performance. Accordingly, we propose the use of the Barlow Twins [30] method to measure the cross-correlation matrix between the corresponding internal outputs of the teacher and student models and ensure it is as close to the identity matrix as possible:

$$C_{j,k} = \frac{\sum_{t=1}^{T_{\text{B}}} \Theta_{t,j}^{\text{CA-SW-SSL}} \Theta_{t,k}^{\text{CSWA}}}{\sqrt{\sum_{t=1}^{T_{\text{B}}} (\Theta_{t,j}^{\text{CA-SW-SSL}})^2} \sqrt{\sum_{t=1}^{T_{\text{B}}} (\Theta_{t,j}^{\text{CSWA}})^2}} \tag{35}$$

$$\mathcal{L}^{\text{TSBT}} = \sum_{j=1}^{C_\Theta} (1 - C_{j,j})^2 + \lambda \sum_{j=1}^{C_\Theta} \sum_{k \neq j} C_{j,k}^2 \tag{36}$$

where C is the cross-correlation matrix computed between the outputs of the teacher and student models along the batch dimension, with values ranging from −1 (perfect anti-correlation) to 1 (perfect correlation); λ is positive constant trading of the importance of the first and second terms of the loss, set to 0.05 based on ablation studies from [30].

Intuitively, the TSBT loss functions by setting the diagonal elements of the cross-correlation matrix to 1 and the off-diagonal elements to 0, ensuring that the embeddings are unaffected by both the teacher and student models. This approach circumvents the challenges linked to optimizing absolute numerical differences and allows for a more nuanced evaluation of each vector component of the embedding, thereby minimizing the risk of local distortions. By substituting MSE-TS with TSBT, the FKD loss becomes:

$$\mathcal{L}^{\text{FKD-TSBT}} = \gamma \mathcal{L}^{\text{TSBT-U}} + (1 - \gamma) \mathcal{L}^{\text{TSBT-V}} \tag{37}$$

where $\mathcal{L}^{\text{TSBT-U}}$ and $\mathcal{L}^{\text{TSBT-V}}$ are the TSBT losses for U, and V, respectively.

5. Experiments and results analysis

5.1. Implementation detail

5.1.1. Data preparation

CA-SW-SSL is first assessed on the DNS-Challenge dataset [54], which provides over 500 hours of clean speech clips spoken by 2150 speakers, and more than 180 hours of noise clips for training. Using the official scripts, we have generated a training set of approximately 3000 h of noisy-clean pairs, with SNR levels ranging from −5 dB to 15 dB. The dataset also includes a non-blind test set for model evaluation, consisting of 150 noisy-clean pairs. Around 300 h of material was split from the training set and used for preliminary validation experiments.

For a broader comparison, the Voicebank+Demand dataset [55] is also considered. This dataset includes 28 speakers in the training set and 2 unseen speakers in the test set. The training set consists of 11,572 noisy-clean pairs, while the test set includes 824 pairs. In the training set, the utterances are corrupted with one of ten noise types, comprising two artificial noises and eight noises from the Demand dataset, across four SNR levels: 0 dB, 5 dB, 10 dB and 15 dB. The test set is generated

Table 1
Hyperparameter configurations for baselines, various versions of CSWA models, and CA-SW-SSL models with different self-supervised learned representations.

Model	N_W, C_W	N_S, C_S	N_{W2S}, C_{W2S}	N_L, C_L	N_G, C_G	N_{S2W}, C_{S2W}	N_{DP}, C_{DP}	SSL	#Param. (M)	MACs. (G/s)
S-TCN	–	4, 512	–	8, 512	–	–	–	–	46.81	2.29
ConvTasnet	15, 128	–	–	–	–	–	–	–	7.8	7.72
CSWA	6, 128	4, 512	8, 64	2, 1024	3, 512	6, 64	–	–	41.73	7.52
CSWA-Lite	4, 128	2, 512	8, 64	1, 1024	1, 512	4, 64	–	–	22.49	4.28
CSWA-Tiny	2, 128	2, 256	6, 64	1, 512	1, 128	2, 64	–	–	6.67	2.80
CA-SW-SSL	6, 128	4, 512	8, 64	2, 1024	3, 512	6, 64	8, 64	wav2vec 2.0	370.64	26.64
								Hubert	369.86	26.58
								WavLM	259.55	21.72

using five unseen noises selected from the Demand dataset, with SNR levels of 2.5 dB, 7.5 dB, 12.5 dB, and 17.5 dB.

Waveforms are sampled at 16 kHz. A 40 ms square-root Hann window is applied for both analysis and synthesis, with a 50% overlap between adjacent frames. This setup results in a 640-point FFT, giving 321 frequency bands.

5.1.2. Evaluation metrics

For the DNS-Challenge dataset, we evaluate the effectiveness of our model using three objective metrics: perceptual evaluation of speech quality (PESQ) [56], short-time objective intelligibility (STOI) [57], and scale-invariant signal-to-noise ratio (SISNR) [58]. PESQ is a widely recognized metric for assessing speech quality, with scores ranging from −0.5 to 4.5, and its wide-band version (WB-PESQ) is employed for evaluation on both datasets. STOI serves as an effective measure of speech intelligibility, with values between 0 and 1. SISNR is commonly used to quantify the level of noise introduced in the estimated speech and overcomes the disadvantage that SNR is susceptible to variations in the energy of the input signal. Three mean opinion score (MOS) related metrics [59], namely CSIG, CBAK, and COVL, are also employed in the Voicebank+Demand experiments. For all these metrics, higher values indicate better speech quality.

5.1.3. Hyperparameter configurations and training

Table 1 summarizes the hyperparameter configurations used in this work, together with their respective parameter counts and multiply-accumulate operations (MACs). The CSWA model configuration is derived from multiple rounds of exploratory trials. Building upon this optimal CSWA setup, we construct CA-SW-SSL with three large-scale SSL models (wav2vec 2.0, HuBERT, and WavLM). Subsequently, CSWA-Lite and CSWA-Tiny are obtained by proportionally reducing layers and channels to approximate the parameter count and MACs of the current SOTA model.

It should be noted that S-TCN [21] (highlighted in purple in Fig. 1) and ConvTasnet [12] (highlighted in pink) serve as the spectral and waveform baselines, respectively. These baselines incorporate more basic modules than their original implementations to ensure comparable parameter counts and MACs with the proposed CSWA.

The training procedure closely follows that of the previous study [60] and the baseline method [12]. Specifically, we train for 100 epochs using the Adam optimizer [61]. The learning rate, initially set to 3×10^{-4} based on a "learning-rate range test" [62], is halved if there is no validation improvement over 3 consecutive epochs, and early stopping is triggered if there is no improvement for 10 epochs. Each experiment conducted 3 independent training runs with distinct random seeds and reported the mean performance of three optimal checkpoints. All experiments were performed on an NVIDIA A100 GPU cluster (4×80GB).[1]

5.2. Performance analysis of CSWA

To validate the effectiveness of our proposed CSWA model, we present a comprehensive comparison of the average WB-PESQ, PESQ, STOI, and SISNR among S-TCN, ConvTasnet and CSWA models on the DNS challenge non-blind test set, as illustrated in Table 2. For convenience, all models were trained using the 300-hour subset. The CSWA model outperforms both the spectrum-based and waveform-based baselines, namely S-TCN and ConvTasnet. Under comparable parameters and computational complexity, the CSWA achieves absolute improvements of 0.97, 0.54, 3.34%, and 5.44 dB in WB-PESQ, PESQ, STOI, and SISNR, respectively, when compared to the best baseline results. This observation demonstrates that the performance gains of CSWA are not merely due to an increase in parameter count or computational load but rather from the innovative structural design that fully leverages the respective advantages of spectrum and waveform modeling.

Further, we conducted ablation experiments on the key CSWA modules, including the W2S cross-attention module, the S2 W cross-attention module and the conformer module. Results for the spectrum,

Table 2
Comparison of average WB-PESQ, PESQ, STOI and SISNR among noisy, baselines, CASW and CA-SW-SSLs with different SSL representations on the DNS challenge non-blind test set. All models were trained using the 300-hour subset.

Model	SSL	WB-PESQ	PESQ	STOI (%)	SISNR
Noisy	–	1.58	2.45	91.52	9.07
S-TCN	–	2.04	2.88	93.72	13.34
ConvTasnet	–	2.09	2.96	94.03	14.21
CSWA	–	**3.06**	**3.50**	**97.37**	**19.65**
CA-SW-SSL	wav2vec 2.0	3.19	3.60	97.73	20.28
	HuBert	3.22	3.61	97.79	20.35
	WavLM	**3.27**	**3.64**	**97.93**	**20.54**

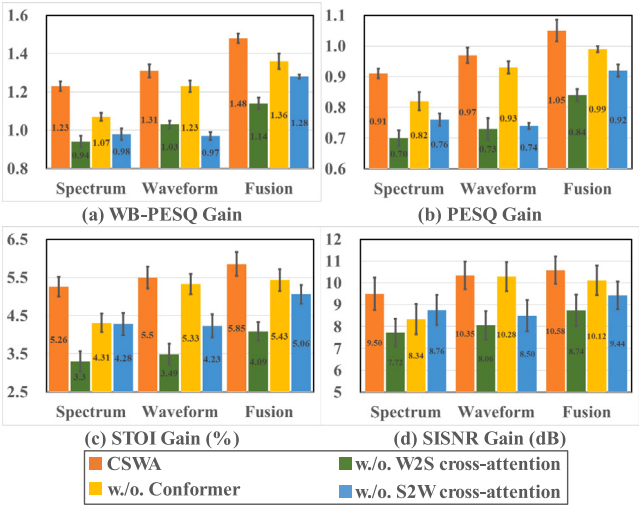

Fig. 3. Comparison of average (a) WB-PESQ gain, (b) PESQ gain, (c) STOI gain and (d) SISNR gain among the CSWA model and its ablation versions on the DNS challenge non-blind test set. Notably, there are three results: one from the spectrum branch, one from the waveform branch, and the fused result. All models were trained using the 300-hour subset.

[1] Source code will be available at *link available upon acceptance.*

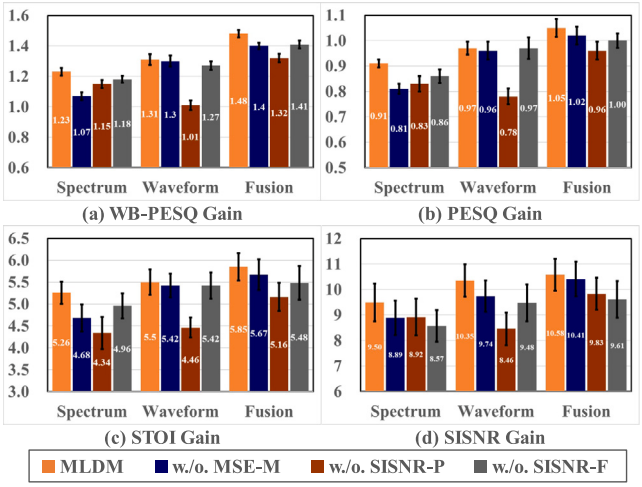

Fig. 4. Comparison of average (a) WB-PESQ gain, (b) PESQ gain, (c) STOI gain and (d) SISNR gain among the MLDM loss and its ablation versions on the DNS challenge non-blind test set. Notably, there are three results: one from the spectrum branch, one from the waveform branch, and the fused result. All models were trained using the 300-hour subset.

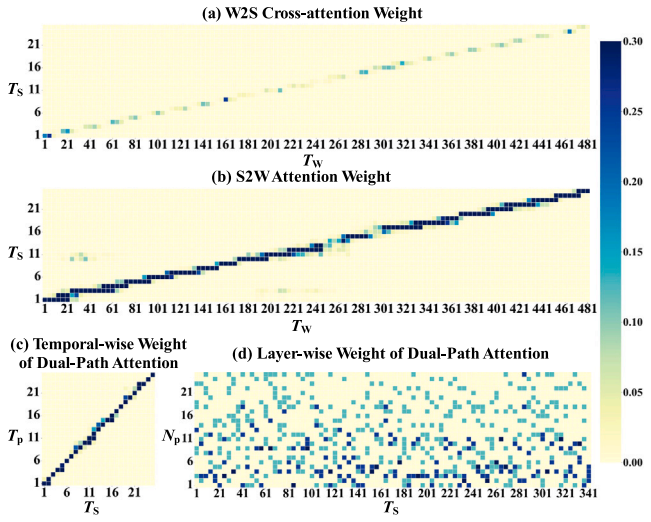

Fig. 5. Visualization of attention weights for the (a) W2S, (b) S2 W, and (c) and (d) dual-path cross-attention modules. Notably, the dual-path cross-attention modules include two types of weights: temporal-wise and layer-wise weights. T_S, T_W and T_P denote the time steps of the spectrum, waveform and SSL representations, respectively. N_P denotes the number of layers in the SSL models.

waveform and fused branches are presented through bar charts with error bars in Fig. 3, illustrating the mean performance metrics and their standard deviations across multiple runs. The relatively small error bars indicate that our findings remain stable under repeated trials. We can observe that as the W2S cross-attention module is ablated, results related to the spectrum branch have a significant degradation, with absolute reductions of 0.29, 0.24, 2.08%, and 2.23 dB in WB-PESQ, PESQ, STOI, and SISNR, respectively. The waveform branch also experiences degradation, with reductions of 0.25, 0.2, 1.91%, and 1.79 dB in WB-PESQ, PESQ, STOI, and SISNR, respectively. However, this degradation is less pronounced than that observed in the spectrum branch. The intuitive explanation is that the spectrum branch depends on the W2S cross-attention module to acquire the missing phase information from the waveform branch, making its performance more reliant on this module. However, the degradation trend is reversed when ablating the S2 W cross-attention module. Specifically, the absolute reductions in WB-PESQ, PESQ, STOI, and SISNR for the waveform branch compared to the spectrum branch are 0.34 vs. 0.25, 0.23 vs. 0.15, 1.27% vs. 0.98%, and 1.85 dB vs. 0.74 dB, respectively. These results can be explained by recalling that the S2 W cross-attention module introduces global context information in the waveform branch, enhancing its performance. Finally, when the Conformer module is ablated, both the spectrum and waveform branches experience a relatively consistent degradation, with absolute reductions of approximately 0.15, 0.1, 0.45%, and 0.5 dB in WB-PESQ, PESQ, STOI, and SISNR, respectively. This observation further underscores the importance of global context information in both spectrum-domain and waveform-domain SE methods.

Next, we carried out a set of ablation experiments to investigate the specific contributions of MSE-M, SISNR-P, and SISNR-F. Results for the spectrum branch, waveform branch, and fused outputs are presented through bar charts with error bars in Fig. 4, illustrating the mean performance metrics and their standard deviations across multiple runs. The relatively small error bars indicate that our findings remain stable under repeated trials. We can observe a consistent degradation trend. Specifically, when MSE-M is ablated, the spectrum branch shows the largest degradation, with absolute reductions of 0.16, 0.10, 0.58%, and 0.61 dB in WB-PESQ, PESQ, STOI, and SISNR, respectively. Conversely, when SISNR-P is ablated, the waveform branch exhibits the most significant degradation, with absolute reductions of 0.3, 0.19, 1.04%, and 1.89 dB in WB-PESQ, PESQ, STOI, and SISNR, respectively. Another interesting observation is that, on the one hand, SISNR-F prevents the fused output from being disproportionately influenced by either branch, experiencing the most severe degradation. On the other hand, without SISNR-F, the degradations from the spectrum and waveform branches accumulate in the fused output. Therefore, we highlight the importance of SISNR-F in intelligently and automatically regulating the contributions of the spectrum and waveform branches, functioning akin to a weight factor but without the need for manual adjustment.

A more thorough understanding of the model's prediction can be gained by visualizing the attention weights for the W2S and S2 W cross-attention modules using heatmaps for a randomly selected utterance from the DNS challenge non-blind test set, as shown in Fig. 5(a) and (b). The clearly visible diagonal patterns indicate a successful temporal alignment between the spectrum and waveform representations, implying that the W2S attention helps supplement missing information, while the S2 W attention extends long-term dependencies from lower to higher temporal resolutions. And we found that the diagonal alignment patterns remain consistent across varied runs and test samples.

5.3. Performance analysis of CA-SW-SSL

In this section, we analyze the CA-SW-SSL model on the DNS-Challenge non-blind test set and compare it with the CSWA model, as shown in Table 2. Three CA-SW-SSL models are built on top of the CSWA model using a different SSL architecture, namely wav2vec2.0-large, Hubert-large, or WavLM-large. CA-SW-SSL models exhibit consistent improvements over the CSWA model, with the CA-SW-SSL-WavLM model achieving the best performance. Specifically, CA-SW-SSL-WavLM attains absolute improvements of 0.21, 0.14, 0.56%, and 0.89 dB in WB-PESQ, PESQ, STOI, and SISNR, respectively, over the CSWA model.

Figs. 5 (c) and (d) visualize attention weights of the DPCA module to gain a better understanding of the observed improvements. DPCA includes two types of weights: temporal- and layer-wise weights. First, Fig. 5(c) demonstrates a clear temporal alignment between the spectrum and SSL representations; moreover, Fig. 5(d) shows another interesting pattern: At all time steps, the lower-layer SSL representations exhibit larger layer-wise weights, indicating that the performance of speech enhancement primarily stems from the abundant acoustic details present in those lower layers rather than the more abstract acoustic knowledge in the higher layers. These patterns are consistent

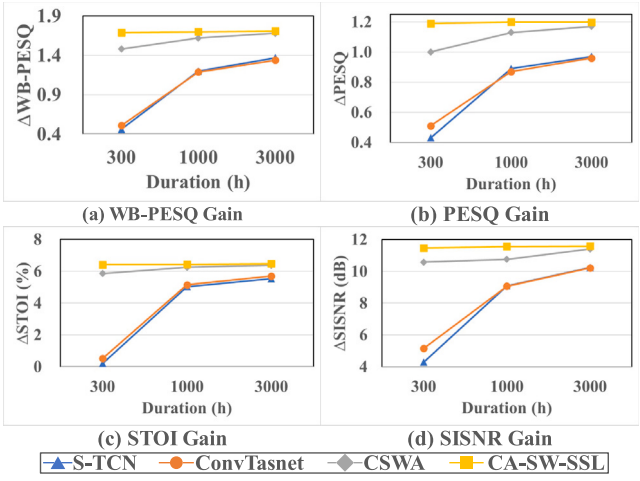

Fig. 6. Comparison of average (a) WB-PESQ gain, PESQ gain, STOI gain and SISNR gain among S-TCN, ConvTasnet, CSWA and CA-SW-SSL model trained with different training set sizes (300 h, 1000 h and 3000 h) on the DNS-Challenge non-blind test set.

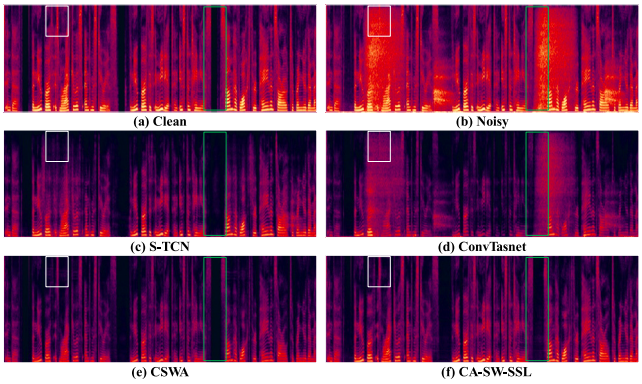

Fig. 7. An utterance example comparing the outputs of different models, including (a) clean, (b) noisy, (c) S-TCN-enhanced, (d) ConvTasnet-enhanced, (e) CSWA-enhanced and (f) CA-SW-SSL-enhanced spectrum features.

across different runs and test samples, providing further evidence of the effectiveness of the CA-SW-SSL model.

We hypothesize that the SSL representations serve a role similar to data augmentation, and we verify our idea via comparative experiments with different sizes of training sets. Results are given in Fig. 6. It can be observed that the performance gap between the CA-SW-SSL and CSWA models gradually narrows as the training set size increases. When the training duration grows from 300h to 3000h, the absolute differences between the CA-SW-SSL and CSWA models in terms of WB-PESQ, PESQ, STOI, and SISNR decrease from 0.21, 0.19, 0.56%, and 0.89 dB to 0.03, 0.03, 0.10%, and 0.18 dB, respectively. This finding provides preliminary evidence that the performance boost from the SSL block diminishes as the amount of training data increases, leading us to think of the SSL module as a data augmentation technique.

Another interesting observation is that the CSWA model demonstrates more consistent performance as the training set size increases than the S-TCN and ConvTasnet models. Specifically, as the training duration increases from 300h to 1000h, the S-TCN and ConvTasnet models exhibit similar significant improvements of approximately 0.7, 0.4, 4.7% and 4.5 dB across WB-PESQ, PESQ, STOI, and SISNR, respectively. In contrast, the CSWA model only shows gains of 0.14, 0.13, 0.4% and 0.18 dB in WB-PESQ, PESQ, STOI, and SISNR, respectively. Hence, the CSWA model effectively uses data via joint spectrum and waveform modeling, demonstrating robustness in low-resource scenarios.

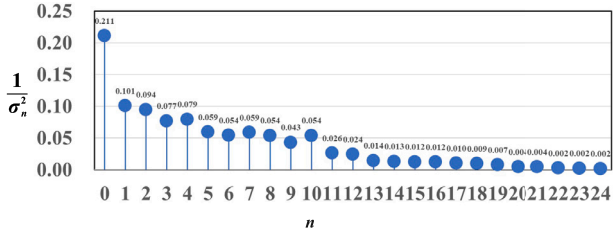

Fig. 8. Visualization of layer-wise distance weights in ALDM. All weights are normalized such that their sum equals 1 for clearer comparison.

Table 3
Comparison of average WB-PESQ, PESQ, STOI (%) and SISNR (dB) between the CA-SW-SSL and CSWA model with different BKD losses on the DNS challenge non-blind test set.

Model	BKD	WB-PESQ	PESQ	STOI(%)	SISNR(dB)
Noisy	–	1.58	2.45	91.52	9.07
CA-SW-SSL	–	3.29	3.65	97.99	20.65
	MSE	3.39	3.69	98.23	21.06
	ALDM	**3.51**	**3.77**	**98.72**	**21.76**
CSWA	–	3.26	3.62	97.89	20.47
	MSE	3.33	3.65	97.97	20.62
	ALDM	**3.40**	**3.71**	**98.13**	**20.95**

Finally, we present a compelling visual comparison illustrating the enhancement results yielded by the S-TCN, ConvTasnet, CSWA, and CA-SW-SSL models in Fig. 7. We utilized a randomly selected example utterance from the DNS challenge non-blind test set, with all spectral features normalized at an utterance level. In the clean spectrum, the white and green boxes highlight activated time-frequency units, which unfortunately become obscured by noise in the noisy spectrum. The S-TCN model eliminates both noise and activated time-frequency units, indicating an issue of over-suppression. Conversely, the ConvTasnet-enhanced spectrum retains both the noise and the activated units, suggesting residual noise exists. Moreover, the CSWA model showcases improved discrimination by removing noise while preserving the original activated time-frequency bands. Notably, the CA-SW-SSL model excels in retaining spectral details, uniquely recovering the weaker activation band on the left side of the white box.

5.4. Performance analysis of BKD

To demonstrate the effectiveness of the proposed BKD framework, we have showcased the WB-PESQ, PESQ, STOI, and SISNR results for the CSWA and CA-SW-SSL models trained with two different BKD losses in Table 3. BKD-MSE enables the CSWA model to achieve comparable performance with the CA-SW-SSL model without increasing the amount of parameters, or the computational load, achieving absolute improvements of 0.04 and 0.005 in WB-PESQ and PESQ, respectively, while maintaining comparable results in STOI and SISNR, specifically 97.97% vs. 97.99% and 20.62 dB vs. 20.65 dB. Moreover, the CA-SW-SSL model also delivered absolute improvements of 0.10, 0.04, 0.24% and 0.41 dB in WB-PESQ, PESQ, STOI, and SISNR, demonstrating the effectiveness and generalizability of the BKD framework in enhancing speech quality and intelligibility across various model backbones.

Furthermore, BKD-ALDM outperforms BKD-MSE, achieving absolute improvements of 0.07, 0.06, 0.16%, and 0.33 dB in WB-PESQ, PESQ, STOI, and SISNR, respectively, in the CSWA case. In the CA-SW-SSL case, BKD-ALDM achieves larger absolute improvements of 0.12, 0.08, 0.49%, and 0.70 dB across the same metrics. These findings demonstrate the importance of distance metrics at different layers, further confirming the effectiveness and generalizability of the BKD framework. In the following sections, we simply use BKD to refer to BKD-ALDM.

Table 4
Comparison of average WB-PESQ, PESQ, STOI (%) and SISNR (dB) across different sizes of CSWA models with various FKD losses on the DNS challenge non-blind test set.

Model	FKD	WB-PESQ	PESQ	STOI(%)	SISNR(dB)
Noisy	–	1.58	2.45	91.52	9.07
CA-SW-SSL+BKD	–	3.51	3.77	98.72	21.76
	-	3.26	3.62	97.89	20.47
CSWA	MSE	3.34	3.64	97.92	20.55
	TSBT	**3.41**	**3.71**	**98.38**	**21.20**
	–	3.40	3.71	98.13	20.95
CSWA+BKD	MSE	3.43	3.72	98.15	21.01
	TSBT	**3.50**	**3.77**	**98.55**	**21.38**
	–	3.16	3.57	97.24	18.72
CSWA-Lite+BKD	MSE	3.21	3.61	97.26	18.89
	TSBT	**3.45**	**3.74**	**98.02**	**20.19**
	–	2.76	3.36	95.95	17.94
CSWA-Tiny+BKD	MSE	2.88	3.43	95.99	17.96
	TSBT	**3.37**	**3.69**	**97.58**	**19.50**

To further elucidate the selectivity of the BKD framework regarding distances across different layers, we analyzed a randomly selected example utterance from the DNS challenge non-blind test set, visualizing its layer-wise distance weights in Fig. 8. Notably, the lower layers exhibited larger weights, consistent with the layer-wise attention weights observed in the dual-path cross-attention visualization. This finding implies that distance at the lower layers, which capture finer acoustic details, plays a pivotal role in enhancing both speech quality and intelligibility.

5.5. Performance analysis of FKD

In the following analysis, we examine the effectiveness of the FKD framework by comparing two FKD losses under the BKD scenario for the CSWA model, as well as its Lite and Tiny variants. As shown in Table 4, the FKD loss narrows the performance gap between the CSWA student and the CA-SW-SSL teacher, while the proposed FKD-TSBT loss surpasses the FKD-MSE baseline. Specifically, with FKD-MSE, the CSWA model improves by 0.08, 0.02, 0.03%, and 0.08 dB in WB-PESQ, PESQ, STOI, and SISNR, respectively. The FKD-TSBT loss achieves even more significant gains of 0.15, 0.09, 0.49%, and 0.73 dB, underscoring the effectiveness of focusing on inter-layer distance correlations rather than absolute values. Furthermore, when combined with BKD, these gains increase to 0.24, 0.15, 0.66%, and 0.91 dB, offering preliminary evidence for the robustness of both FKD and BiKD. For brevity, we use FKD to refer to FKD-TSBT in the subsequent discussion.

Another noteworthy observation is that the improvements increase as the complexity of the student models decreases. Concretely, CSWA-Lite reduces 19.24 M parameters and 3.24 G/s relative to CSWA yet still achieves gains of 0.24, 0.13, 0.76%, and 1.30 dB in the metrics above. CSWA-Tiny trims an additional 15.82 M parameters and 1.48 G/s, yielding even larger improvements of 0.49, 0.26, 1.59%, and 1.54 dB. These trends validate the adaptability of FKD across different model sizes and especially highlight substantial gains for lower-complexity variants. Although these lightweight models remain slightly behind CA-SW-SSL in overall accuracy, the substantial reductions in parameter count and computational overhead justify a moderate trade-off in performance for many resource-limited applications.

To further analyze the impact of the FKD framework on model predictions, we randomly selected an example utterance from the DNS Challenge non-blind test set and conducted a T-SNE visualization of frame-level hidden representations, denoted as U_t and V_t. As illustrated in Fig. 9, without FKD, the hidden representations from the CSWA+BKD model, i.e., U_t^{CSWA} and V_t^{CSWA} appear more dispersed and exhibit a less coherent structure. Notably, some predictions deviate from the high-performance subspace enveloped by the hidden representations of the CA-SW-SSL+BKD model, i.e., $U_t^{\text{CA-SW-SSL}}$ and $V_t^{\text{CA-SW-SSL}}$. However, after applying FKD, both U_t^{CSWA} and V_t^{CSWA} are effectively aligned within this high-performance subspace.

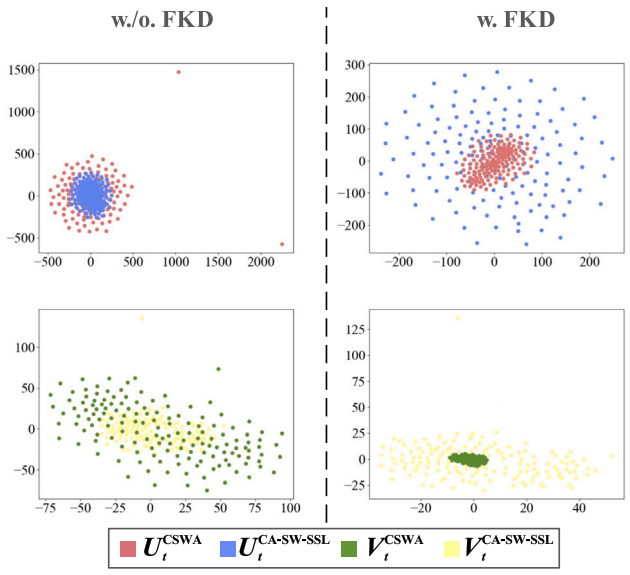

Fig. 9. Comparative t-SNE analysis of frame-level representations in CA-SW-SSL and CSWA models: With and Without FKD.

5.6. Comparison with SOTA

Table 5 presents a thorough comparison between our method and prior SOTA methods on the DNS Challenge non-blind test set. Our CSWA-Lite+BiKD model has demonstrated remarkable improvements in WB-PESQ, PESQ, and STOI metrics compared to the previous SOTA TaEr model, with absolute improvements of 0.19, 0.14, 0.46% and 0.79 dB. These improvements were achieved while maintaining comparable computational complexity despite a higher parameter count. Furthermore, our CSWA-Lite+BiKD model has showcased still significant absolute gains of 0.11, 0.09, 0.02% and 0.10 dB in WB-PESQ, PESQ, and STOI, with similar parameter counts and only half the computational load. These experimental results emphasize the superiority and efficiency of our proposed method.

However, most recent SOTA methods for the DNS-Challenge non-blind test set essentially utilize spectrum-domain models. To support the comparison with similar methods and showcase the generalization capability of our methods, we conducted supplementary experiments using the Voicebank+Demand dataset. In Table 6, we present a comparison of average WB-PESQ, CSIG, CBAK and COVL among several well-known joint spectrum-waveform methods, SSL-based methods and our methods. It is evident that our CA-SW-SSL+BKD model outperforms the top-performing SSL-based method, yielding absolute improvements of 0.06, 0.05, 0.44 and 0.01 in WB-PESQ, CSIG, CBAK, and COVL, respectively. Moreover, our CSWA-Tiny+BiKD model outperforms the top-performing joint spectrum-waveform methods, demonstrating absolute improvements of 0.05, 0.08, 0.08 and 0.07 in WB-PESQ, CSIG, CBAK, and COVL, respectively. These results not only show the superiority of our methods over similar ones but also underline its generalization across diverse datasets.

6. Conclusion

This work presented an innovative CA-SW-SSL model that integrates spectrum, waveform, and SSL features for SE. Experimental findings highlight that cross-domain information exchange between spectrum and waveform branches significantly improves individual and fused performance. The performance gains from using SSL features as inputs, similar to data augmentation, diminish as the training data size increases. Moreover, lower-level SSL features capture essential acoustic details crucial for SE. We have also introduced a novel BiKD

Table 5
Comparison of average WB-PESQ, PESQ, STOI (%) and SISNR (dB) between previous SOTA methods and our methods on the DNS challenge non-blind test set.

Model	Features	#Param. (M)	MACs (G/s)	WB-PESQ	PESQ	STOI (%)	SISNR (dB)
NSNet [63]	Mag	1.26	–	2.15	2.87	94.47	15.61
DTLN [64]	Mag	0.99	–	–	3.04	94.76	16.34
DCCRN [65]	Real+Imag	3.67	14.06	–	3.27	–	–
PoCoNet [66]	Real+Imag	50.00	–	2.75	–	–	–
FullSubNet [67]	Mag	5.64	14.92	2.78	3.31	96.11	17.29
TRU-Net [68]	Mag	0.38	–	2.86	3.36	96.32	17.55
DCCRN+ [69]	Real+Imag	3.30	–	–	3.33	–	–
CTS-Net [70]	Mag+Real+Imag	4.35	5.57	2.94	3.42	96.66	17.99
GaGNet [71]	Real+Imag	6.01	1.64	3.17	3.56	97.13	18.91
FRNet [72]	Real+Imag	7.52	2.81	3.14	3.52	96.91	18.75
FullSubNet+ [73]	Mag+Real+Imag	8.67	–	2.98	3.50	96.69	18.34
FS-CANet [74]	Mag	4.21	–	3.02	3.51	96.74	18.08
FRCRN [75]	Real+Imag	10.27	241.98	3.23	3.60	97.69	19.78
STSubNet [76]	Mag	5.66	–	3.00	–	97.03	19.64
TaEr [77]	Real+Imag	6.42	4.36	3.26	3.60	97.56	19.40
CA-SW-SSL+BKD	Mag+Wave+SSL	259.55	21.72	3.51	3.77	98.72	21.76
CSWA+BiKD	Mag+Wave	41.73	7.52	3.50	3.77	98.55	21.38
CSWA-Lite+BiKD	**Mag+Wave**	**22.49**	**4.28**	**3.45**	**3.74**	**98.02**	**20.19**
CSWA-Tiny+BiKD	**Mag+Wave**	**6.67**	**2.80**	**3.37**	**3.69**	**97.58**	**19.50**

Table 6
Comparison of average WB-PESQ, CSIG, CBAK and COVL among several well-known joint spectrum-waveform methods, SSL-based methods and our methods on the Voicebank+Demand test set.

Model	Features	WB-PESQ	CSIG	CBAK	COVL
Noisy	–	1.97	3.35	2.44	2.63
TFT-Net [14]	Mag+Wave	2.75	3.93	3.44	3.34
MDPhD [13]	Mag+Wave	2.70	3.85	3.39	3.27
WMPNet [33]	Mag+Wave	3.05	4.27	3.53	3.68
WSFNet [34]	**Real+Imag+Wave**	**3.09**	**4.32**	**3.51**	**3.72**
[18]	SSL	2.80	–	–	–
[19]	Mag+SSL	3.20	4.53	3.60	3.88
[41]	Mag+SSL	2.79	4.10	2.68	3.44
PFPL [38]	**Real+Imag+SSL**	3.15	4.18	3.60	3.67
PCS-CS-WavLM [20]	Mag+SSL	**3.54**	**4.75**	**3.54**	**4.25**
CA-SW-SSL+BKD	**Mag+Wave+SSL**	**3.60**	**4.80**	**3.98**	**4.26**
CSWA+BiKD	Mag+Wave	3.46	4.63	3.83	4.09
CSWA-Lite+BiKD	Mag+Wave	3.27	4.47	3.70	3.90
CSWA-Tiny+BiKD	**Mag+Wave**	**3.14**	**4.40**	**3.59**	**3.79**

framework to address the increased parameter count and computational complexity caused by the SSL model. Results have illustrated the strong correlation between the distance measure in the lower-level SSL representation space and speech quality. Furthermore, optimizing the cross-correlation matrix of intermediate-layer features has enhanced knowledge transfer to smaller models more effectively than optimizing numerical differences. In future work, we plan to expand the CA-SW-SSL model by incorporating additional data sources, e.g., video, and explore pruning techniques to reduce model complexity further.

Although the model performs well on the tested datasets, challenges remain when encountering more diverse noise conditions and unseen languages. Future work will thus focus on extending CA-SW-SSL to handle broader data modalities (e.g., multimodal inputs such as video) and exploring pruning or quantization techniques to further reduce its computational footprint.

CRediT authorship contribution statement

Hang Chen: Writing – original draft, Methodology, Formal analysis, Conceptualization. **Chenxi Wang:** Validation, Software. **Qing Wang:** Writing – review & editing. **Jun Du:** Supervision, Funding acquisition. **Sabato Marco Siniscalchi:** Writing – review & editing, Conceptualization. **Genshun Wan:** Resources. **Jia Pan:** Investigation. **Huijun Ding:** Writing – review & editing, Visualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data used in this article is a public dataset and the code will be released after acceptance.

References

[1] P.C. Loizou, Speech Enhancement: Theory and Practice, CRC, 2007.
[2] S. Boll, Suppression of acoustic noise in speech using spectral subtraction, IEEE Trans. Acoust. Speech Signal Process. 27 (2) (1979) 113–120.
[3] Y. Ephraim, D. Malah, Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator, IEEE Trans. Acoust. Speech Signal Process. 32 (6) (1984) 1109–1121.
[4] I. Cohen, Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging, IEEE Trans. Acoust. Speech Signal Process. (2003) 466–475.
[5] Y. Xu, J. Du, L.R. Dai, C.H. Lee, A regression approach to speech enhancement based on deep neural networks, IEEE/ ACM Trans. Audio Speech Lang. Process. 23 (1) (2014) 7–19.
[6] Y. Wang, D.L. Wang, Towards scaling up classification-based speech separation, IEEE/ ACM Trans. Audio Speech Lang. Process. 21 (7) (2013) 1381–1390.
[7] Y. Wang, A. Narayanan, D.L. Wang, On training targets for supervised speech separation, IEEE/ ACM Trans. Audio Speech Lang. Process. 22 (12) (2014) 1849–1858.
[8] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.
[9] Y. Wang, A. Narayanan, D. Wang, On training targets for supervised speech separation, IEEE/ ACM Trans. Audio Speech Lang. Process. 22 (12) (2014) 1849–1858.
[10] Z.Q. Wang, G. Wichern, J. Le Roux, On the compensation between magnitude and phase in speech separation, IEEE Signal Process. Lett. 28 (2021) 2018–2022.
[11] S. Pascual, A. Bonafonte, J. Serra, SEGAN: Speech enhancement generative adversarial network, in: Proc. Interspeech 2017, 2017, pp. 3642–3646.
[12] Y. Luo, N. Mesgarani, Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation, IEEE/ ACM Trans. Audio Speech Lang. Process. 27 (8) (2019) 1256–1266.
[13] J.H. Kim, J. Yoo, S. Chun, et al., Multi-domain processing via hybrid denoising networks for speech enhancement, 2018, arXiv preprint arXiv:1812.08914.
[14] C. Tang, C. Luo, Z. Zhao, et al., Joint time-frequency and time domain learning for speech enhancement, in: Proc. IJCAI 2021, 2021, pp. 3816–3822.
[15] K. Zhang, S. He, H. Li, et al., DBNet: A dual-branch network architecture processing on spectrum and waveform for single-channel speech enhancement, in: Proc. Interspeech 2021, 2021.
[16] A. Baevski, A. Mohamed, Effectiveness of self-supervised pre-training for asr, in: Proc. ICASSP 2020, 2020, pp. 7694–7698.

[17] C.I. Lai, Y.S. Chuang, H.Y. Lee, et al., Semi-supervised spoken language understanding via self-supervised speech and language model pretraining, in: Proc. ICASSP 2021, 2021, pp. 7468–7472.

[18] Z. Huang, S. Watanabe, S.w. Yang, et al., Investigating self-supervised learning for speech enhancement and separation, in: Proc. ICASSP 2022, 2022, pp. 6837–6841.

[19] K.H. Hung, S. wei Fu, H.H. Tseng, et al., Boosting self-supervised embeddings for speech enhancement, in: Proc. Interspeech 2022, 2022, pp. 186–190.

[20] M.S. Khan, M.L. Quatra, K.H. Hung, et al., Exploiting consistency-preserving loss and perceptual contrast stretching to boost SSL-based speech enhancement, in: Proc. MMSP 2024, 2024, pp. 1–6.

[21] A. Li, W. Liu, X. Luo, et al., ICASSP 2021 deep noise suppression challenge: Decoupling magnitude and phase optimization with a two-stage deep network, in: Proc. ICASSP 2021, 2021, pp. 6628–6632.

[22] C. Buciluǎ, R. Caruana, A. Niculescu-Mizil, Model compression, in: Proc. SIGKDD 2006, 2006, pp. 535–541.

[23] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, in: Proc. NIPS 2014 Deep Learning Workshop, 2015.

[24] J. Ba, R. Caruana, Do deep nets really need to be deep? Adv. Neural Inf. Process. Syst. 27 (2014).

[25] G. Urban, K.J. Geras, S.E. Kahou, et al., Do deep convolutional nets really need to be deep and convolutional? in: Proc. ICLR 2017, 2017.

[26] W. Shin, H.J. Park, J.S. Kim, et al., Multi-View Attention Transfer for Efficient Speech Enhancement, in: Proc. Interspeech 2022, 2022, pp. 1198–1202.

[27] J. Cheng, R. Liang, Y. Xie, et al., Cross-Layer Similarity Knowledge Distillation for Speech Enhancement, in: Proc. Interspeech 2022, 2022, pp. 926–930.

[28] Y. Wan, Y. Zhou, X. Peng, et al., ABC-KD: Attention-Based-Compression Knowledge Distillation for Deep Learning-Based Noise Suppression, in: Proc. INTERSPEECH 2023, 2023, pp. 2528–2532.

[29] A. Kendall, Y. Gal, What uncertainties do we need in bayesian deep learning for computer vision? Adv. Neural Inf. Process. Syst. 30 (2017).

[30] J. Zbontar, L. Jing, I. Misra, et al., Barlow twins: Self-supervised learning via redundancy reduction, in: Proc. ICML 2021, 2021, pp. 12310–12320.

[31] D. Wang, J. Lim, The unimportance of phase in speech enhancement, IEEE Trans. Acoust. Speech Signal Process. 30 (4) (1982) 679–681.

[32] J. Lee, H.G. Kang, Real-time neural speech enhancement based on temporal refinement network and channel-wise gating methods, Digit. Signal Process. 133 (2023) 103879.

[33] X. Xiang, X. Zhang, Joint waveform and magnitude processing for monaural speech enhancement, Appl. Acoust. 200 (2022) 109077.

[34] R. Yu, W. Chen, Z. Ye, A novel target decoupling framework based on waveform-spectrum fusion network for monaural speech enhancement, Digit. Signal Process. 141 (2023) 104150.

[35] M. Kolbæk, D. Yu, Z.H. Tan, et al., Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks, IEEE/ ACM Trans. Audio Speech Lang. Process. 25 (10) (2017) 1901–1913.

[36] S.W. Fu, C.F. Liao, T.A. Hsieh, et al., Boosting objective scores of a speech enhancement model by metricgan post-processing, in: Proc. APSIPA ASC 2020, IEEE, 2020, pp. 455–459.

[37] F.G. Germain, Q. Chen, V. Koltun, Speech denoising with deep feature losses, in: Proc. Interspeech 2019, 2019, pp. 2723–2727.

[38] T.A. Hsieh, C. Yu, S.W. Fu, et al., Improving perceptual quality by phone-fortified perceptual loss using wasserstein distance for speech enhancement, in: Proc. Interspeech 2021, 2021, pp. 196–200.

[39] I. Olkin, F. Pukelsheim, The distance between two random vectors with given dispersion matrices, Linear Algebra Appl. (ISSN: 0024-3795) 48 (1982) 257–263.

[40] S. Schneider, A. Baevski, R. Collobert, et al., wav2vec: Unsupervised pre-training for speech recognition, in: Proc. Interspeech 2019, 2019, pp. 3465–3469.

[41] G. Close, W. Ravenscroft, T. Hain, S. Goetze, Perceive and predict: Self-supervised speech representation based loss functions for speech enhancement, in: Proc. ICASSP 2023, 2023, pp. 1–5.

[42] X. Xu, C. Han, Y. Zhang, et al., Curricular contrastive regularization for speech enhancement with self-supervised representations, in: Proc. ICASSP 2024, 2024, pp. 10486–10490.

[43] J. Kim, S. Park, N. Kwak, Paraphrasing complex network: Network compression via factor transfer, Adv. Neural Inf. Process. Syst. 31 (2018).

[44] S.I. Mirzadeh, M. Farajtabar, A. Li, et al., Improved knowledge distillation via teacher assistant, in: Proc. AAAI 2020, vol. 34, (04) 2020, pp. 5191–5198.

[45] A. Romero, N. Ballas, S.E. Kahou, et al., Fitnets: Hints for thin deep nets, in: Proc. ICLR 2015, 2015.

[46] S. Zagoruyko, N. Komodakis, Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer, in: Proc. ICLR 2017, 2017.

[47] D. Chen, J.P. Mei, Y. Zhang, et al., Cross-layer distillation with semantic calibration, in: Proc. AAAI 2021, vol. 35, (8) 2021, pp. 7028–7036.

[48] G. Zhou, Y. Fan, R. Cui, et al., Rocket launching: A universal and efficient framework for training well-performing light net, in: Proc. AAAI 2018, vol. 32, (1) 2018.

[49] B. Heo, M. Lee, S. Yun, et al., Knowledge transfer via distillation of activation boundaries formed by hidden neurons, in: Proc. AAAI 2019, vol. 33, (01) 2019, pp. 3779–3787.

[50] R.D. Nathoo, M. Kegler, M. Stamenovic, Two-step knowledge distillation for tiny speech enhancement, in: Proc. ICASSP 2024, 2024, pp. 10141–10145.

[51] A. Gulati, J. Qin, C.-C. Chiu, et al., Conformer: Convolution-augmented transformer for speech recognition, in: Proc. Interspeech 2020, 2020, pp. 5036–5040.

[52] H. Chen, Q. Wang, J. Du, B.C. Yin, J. Pan, C.H. Lee, Optimizing audio-visual speech enhancement using multi-level distortion measures for audio-visual speech recognition, IEEE/ ACM Trans. Audio Speech Lang. Process. 32 (2024) 2508–2521.

[53] D.S. Williamson, Y. Wang, D. Wang, Complex ratio masking for monaural speech separation, IEEE/ ACM Trans. Audio Speech Lang. Process. 24 (3) (2015) 483–492.

[54] C.K. Reddy, V. Gopal, R. Cutler, et al., The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results, in: Proc. Interspeech 2020, 2020, pp. 2492–2496.

[55] C. Valentini-Botinhao, X. Wang, S. Takaki, et al., Investigating RNN-based speech enhancement methods for noise-robust text-to-speech, in: Proc. ISCA SSW 2016, 2016, pp. 146–152.

[56] A. Rix, J. Beerends, M. Hollier, et al., Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs, in: Proc. ICASSP 2001, vol. 2, 2001, pp. 749–752.

[57] C.H. Taal, R.C. Hendriks, R. Heusdens, et al., An algorithm for intelligibility prediction of time–frequency weighted noisy speech, IEEE Trans. Audio Speech Lang. Process. 19 (7) (2011) 2125–2136.

[58] J.L. Roux, S. Wisdom, H. Erdogan, et al., SDR - half-baked or well done? in: Proc. ICASSP 2019, 2019, pp. 626–630.

[59] Y. Hu, P.C. Loizou, Evaluation of objective quality measures for speech enhancement, IEEE Trans. Audio Speech Lang. Process. 16 (1) (2008) 229–238.

[60] H. Chen, J. Du, Y. Hu, et al., Correlating subword articulation with lip shapes for embedding aware audio-visual speech enhancement, Neural Netw. 143 (2021) 171–182.

[61] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Proc. ICLR 2015, 2015, pp. 1–15.

[62] L.N. Smith, Cyclical learning rates for training neural networks, in: 2017 IEEE Winter Conference on Applications of Computer Vision, WACV, 2017, pp. 464–472.

[63] Y. Xia, S. Braun, C.K. Reddy, et al., Weighted speech distortion losses for neural-network-based real-time speech enhancement, in: Proc. ICASSP 2020, 2020, pp. 871–875.

[64] N.L. Westhausen, B.T. Meyer, Dual-signal transformation LSTM network for real-time noise suppression, in: Proc. Interspeech 2020, 2020, pp. 2477–2481.

[65] Y. Hu, Y. Liu, S. Lv, et al., DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement, in: Proc. Interspeech 2020, 2020, pp. 2472–2476.

[66] U. Isik, R. Giri, N. Phansalkar, et al., PoCoNet: Better speech enhancement with frequency-positional embeddings, semi-supervised conversational data, and biased loss, in: Proc. Interspeech 2020, 2020, pp. 2487–2491.

[67] X. Hao, X. Su, R. Horaud, et al., Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement, in: Proc. ICASSP 2021, 2021, pp. 6633–6637.

[68] H.S. Choi, S. Park, J.H. Lee, et al., Real-time denoising and dereverberation wtih tiny recurrent U-net, in: Proc. ICASSP 2021, 2021, pp. 5789–5793.

[69] S. Lv, Y. Hu, S. Zhang, et al., DCCRN+: Channel-wise subband DCCRN with SNR estimation for speech enhancement, in: Proc. Interspeech 2021, 2021, pp. 2816–2820.

[70] A. Li, W. Liu, C. Zheng, et al., Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement, IEEE/ ACM Trans. Audio Speech Lang. Process. 29 (2021) 1829–1843.

[71] A. Li, C. Zheng, L. Zhang, et al., Glance and gaze: A collaborative learning framework for single-channel speech enhancement, Appl. Acoust. (ISSN: 0003-682X) 187 (2022) 108499.

[72] A. Li, C. Zheng, G. Yu, et al., Filtering and refining: A collaborative-style framework for single-channel speech enhancement, IEEE/ ACM Trans. Audio Speech Lang. Process. 30 (2022) 2156–2172.

[73] J. Chen, Z. Wang, D. Tuo, et al., FullSubNet+: Channel attention fullsubnet with complex spectrograms for speech enhancement, in: Proc. ICASSP 2022, 2022, pp. 7857–7861.

[74] J. Chen, W. Rao, Z. Wang, et al., Speech enhancement with fullband-subband cross-attention network, in: Proc. Interspeech 2022, 2022, pp. 976–980.

[75] S. Zhao, B. Ma, K.N. Watcharasupat, et al., FRCRN: Boosting feature representation using frequency recurrence for monaural speech enhancement, in: Proc. ICASSP 2022, 2022, pp. 9281–9285.

[76] F. Xiong, W. Chen, P. Wang, et al., Spectro-temporal SubNet for real-time monaural speech denoising and dereverberation, in: Proc. Interspeech 2022, 2022, pp. 931–935.

[77] A. Li, G. Yu, C. Zheng, et al., A general unfolding speech enhancement method motivated by taylor's theorem, IEEE/ ACM Trans. Audio Speech Lang. Process. 31 (2023) 3629–3646.