EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

TFDense-GAN: a generative adversarial network for single-channel speech enhancement

Haoxiang Chen[1], Jinxiu Zhang[1], Yaogang Fu[1], Xintong Zhou[1], Ruilong Wang[1*], Yanyan Xu[1] and Dengfeng Ke[2]

*Correspondence: wruilong@bjfu.edu.cn

[1] Beijing Forestry University, 35 Qinghua East Road, Haidian District, Beijing 100083, China
[2] Beijing Language and Culture University, 15 Xueyuan Road, Haidian District, Beijing 100083, China

Abstract

Research indicates that utilizing the spectrum in the time–frequency domain plays a crucial role in speech enhancement tasks, as it can better extract audio features and reduce computational consumption. For the speech enhancement methods in the time–frequency domain, the introduction of attention mechanisms and the application of DenseBlock have yielded promising results. In particular, the Unet architecture, which comprises three main components, the encoder, the decoder, and the bottleneck, employs DenseBlock in both the encoder and the decoder to achieve powerful feature fusion capabilities with fewer parameters. In this paper, in order to enhance the advantages of the aforementioned methods for speech enhancement, we propose a Unet-based time–frequency domain denoising model called TFDense-Net. It utilizes our improved DenseBlock for feature extraction in both the encoder and the decoder and employs an attention mechanism in the bottleneck for feature fusion and denoising. The model has demonstrated excellent performance for speech enhancement tasks, achieving significant improvements in the Si-SDR metric compared to other state-of-the-art models. Additionally, to further enhance the denoising performance and increase the receptive field of the model, we introduce a multi-spectrogram discriminator based on multiple STFTs. Since the discriminator loss can observe the correlations between spectra that traditional loss functions cannot detect, we train TFDense-Net as a generator against the multi-spectrogram discriminator, resulting in a significant improvement in the denoising performance, and we name this enhanced model TFDense-GAN. We evaluate our proposed TFDense-Net and TFDense-GAN on two public datasets: the VCTK + DEMAND dataset and the Interspeech Deep Noise Suppression Challenge dataset. Experimental results show that TFDense-GAN outperforms most existing models in terms of STOI, PESQ, and Si-SDR, achieving state-of-the-art results. The comparison samples of TFDense-GAN and other models can be accessed from https://github.com/yhsjoker/TFDense-GAN.

Keywords: Speech enhancement, Time–frequency domain, Generative adversarial network, Improved DenseBlock, Time–frequency transformer

1 Introduction

Speech enhancement is one of the most extensively studied topics in the field of speech processing. Speech enhancement models take noisy speech signals from challenging environments as input, and their goal is to map the noisy speech signals to clean ones. Speech enhancement models are frequently employed to reduce background noise, echoes, and other interferences. As a result, they significantly improve the intelligibility and quality of speech communication.

Speech enhancement techniques are generally classified into time domain methods and time–frequency domain methods. Time domain methods directly process the amplitude and temporal features of speech signals within the time domain, taking one-dimensional mixture signal as input. These approaches are straightforward to implement and capable of preserving complete audio information, resulting in clear and explicit audio processing [1, 2]. For instance, Conv-TasNet utilizes temporal convolutional networks (TCNs), composed of stacked one-dimensional dilated convolution blocks, enabling the network to capture the long-term dependencies of speech signals while maintaining a compact model size [3]. Dang et al. introduce the CleanUnet model, which employs a Unet architecture to progressively downsample the one-dimensional time series, extracting feature vectors and utilizing a transformer encoder in the bottleneck for feature fusion and denoising [4]. However, due to characteristics such as high sampling rates of audio sequences and less pronounced audio features, these models do not perform well when the scale of input parameters is significantly large.

Based on signal theory, time–frequency domain methods are extensively used in audio signal processing tasks. These methods typically employ the short-time Fourier transform (STFT) to convert one-dimensional signal into a two-dimensional spectrum. By utilizing a sliding window for Fourier transformation, STFT shortens the length of the audio signal sequence and extracts neighboring features of the audio signal sequence, thereby enhancing the models' abilities for speech tasks. Time–frequency domain speech enhancement techniques aim to transform speech signals into spectrograms via STFT, and then, after noise reduction processing, reconstruct the signals back to the time domain through inverse transformation. Time–frequency domain methods are compact and demonstrate good performance in complex environments and signal distortion scenarios due to their rich audio characteristics, making them widely applied in speech enhancement studies.

Regarding time–frequency network architectures, Kim et al. introduce a novel transformer variant named transformer with Gaussian-weighted self-attention, where the attention weights are adjusted based on the distance between the target and contextual tokens [5]. This mechanism applies a Gaussian weight matrix element-wise to the score matrix to adjust the attention weights. The elements of the Gaussian weight matrix are trainable parameters, allowing the model to learn contextual localization based on the training data. This innovative approach enhances the model's ability to focus on relevant features dynamically, tailored to each specific context. Besides, Dang et al. propose a new dual-path transformer model to implement the feature fusion of two-dimensional spectra, which employs the DenseBlock structure for encoding and decoding, demonstrating that DenseBlock can enhance the performance of speech enhancement models [6]. Moreover, Zhao et al. show that using the complex channel-attention module [7], the complex spatial-attention module, and other technologies to process the spectrum can improve the performance of speech enhancement models [8].

With the progress in artificial intelligence-generated content (AIGC), generative adversarial networks (GANs) have been increasingly utilized in various speech signal processing tasks. The Mel spectrum, obtained through filtering transformations, more closely aligns with human auditory perception but lacks phase information, making the conversion from Mel spectrum to speech inherently lossy. As a result, DNN-based vocoder technologies have been developed for speech synthesis to perform Mel-to-audio conversions. In the field of speech enhancement, due to the irreversibility of Mel filtering, Liu et al. propose a vocoder model named T-F GAN, which maps the denoised Mel spectrum to output speech signals [9].

Among existing speech enhancement methods, the DenseBlock structure and the dual-path transformer have demonstrated high performance in spectrum denoising. However, prior to using the transformer, DPT-FSNet [6] does not incorporate the Unet architecture for downsampling, which increases computational consumption and fails to utilize the advantages of Unet for feature fusion. Also, the dual-path transformer is tasked with calculating the attention for each full-band and sub-band, which presents challenges in extracting features from the overall spectrum.

Therefore, to leverage the advantages of Unet for feature extraction, we employ the encoder in Unet for downsampling before performing feature fusion with attention mechanisms and utilize the decoder for upsampling. During the fusion stage, to extract attention from both full-band and sub-band features, we introduce a global transformer to extract global features. For the bottleneck, we optimize the transformer by integrating three different approaches and get our time–frequency transformer.

Additionally, inspired by the vocoder technology and based on the traditional audio reconstruction loss, we design a multi-spectrogram discriminator to evaluate our generator module and backpropagate the evaluation loss, further enhancing the audio quality. This approach not only leverages the power of GANs but also enhances the practical application of deep learning for improving speech quality and intelligibility across various acoustic environments.

Benefiting from the proposed multi-spectrogram discriminator, which provides adversarial loss for the speech enhancement model during GAN training, the denoising model can focus on the underlying correlations within the speech signal rather than just the one-to-one correspondence between spectral values. Therefore, we combine the discriminator with the generator to improve denoising performance, enhancing the model's robustness and efficiency in various environments.

To summarize, this study has the following contributions.

1. We develop a supervised speech enhancement model based on spectral masking, named TFDense-Net. By optimizing the dilated DenseBlock, we fully take advantage of its feature fusion capabilities, improving the efficiency of upsampling and downsampling in both the encoder and the decoder. For the bottleneck, we utilize our proposed time–frequency transformer to perform feature fusion and denoising. TFDense-Net is evaluated on the VCTK + DEMAND dataset and the Interspeech Deep Noise

Suppression Challenge dataset. Experimental results demonstrate that it has exceptional noise reduction capabilities.

2. Inspired by the widespread application of GANs, we introduce the multi-spectrogram discriminator. This module takes denoised speech signals and clean speech signals as inputs, utilizing STFT to transform the signals into multiple spectra with varying window sizes and hop lengths. The features are subsequently fused through convolution layers, followed by multi-scale and multi-period discriminators to score every 16 frames of audio. Our experiments confirm that the multi-spectrogram discriminator can effectively improve the noise reduction capabilities.

3. To achieve more effective noise reduction in the spectral representation, we incorporate the proposed multi-spectrogram discriminator into TFDense-Net, resulting in a GAN-based model named TFDense-GAN for speech enhancement tasks. We conduct denoising experiments with this model under various signal-to-noise ratios (SNRs) and statistically analyze the evaluation metrics. The experimental results demonstrate that TFDense-GAN exhibits state-of-the-art performance in terms of PESQ and STOI metrics on both the VCTK + DEMAND dataset and the Interspeech Deep Noise Suppression Challenge dataset, compared to other state-of-the-art models.

The remaining sections of this paper are organized as follows. Section 2 introduces the fundamental models and studies related to this work. In Section 3, we detail our proposed time–frequency transformer, TFDense-Net, multi-spectrogram discriminator and TFDense-GAN. Section 4 describes the datasets and evaluation metrics used in this paper. The experimental results are illustrated and discussed in Section 5. Finally, we conclude the paper in Section 6.

2 Related work

2.1 Multi-head attention

In this paper, we employ the multi-head attention [10, 11] to divide the hidden state fusions into multiple heads and form multiple sub-semantic spaces, which allows the model to focus on information from different dimensional spaces, effectively solving the feature fusion problem. The multi-head attention model can be formalized as follows:

$$Q_i, K_i, V_i = XW_i^q, XW_i^K, XW_i^v, i \in [1, H] \tag{1}$$

$$h_i = \text{Attention}(Q_i, K_i, V_i) = \text{SoftMax}\left(\frac{Q_i K_i^T}{\sqrt{d}}\right) V_i \tag{2}$$

$$\text{Mid} = \text{LayerNorm}(X + \text{Concat}(h_1, \ldots, h_i) W_O) \tag{3}$$

where $X \in \mathbb{R}^{l \times d}$ is the input sequences with length l and dimension d, and $Q_i, K_i, V_i \in \mathbb{R}^{l \times d/h}$ are the mapped queries, keys and values, respectively. $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d \times d/h}$ and $W^O \in \mathbb{R}^{d \times d}$ are linear transformation matrices, and Mid is the output from the multi-head attention block.

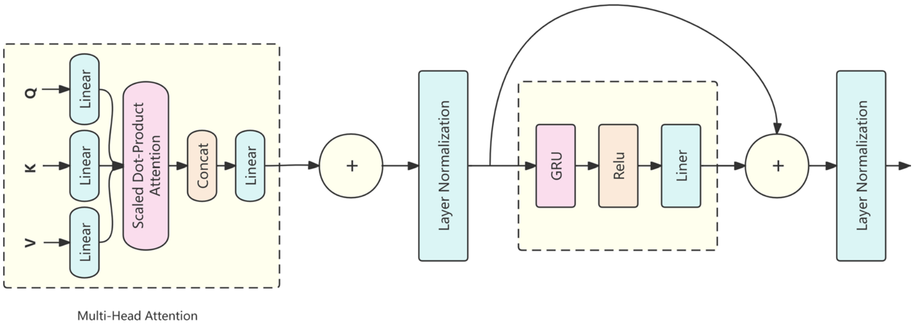


Fig. 1 The improved transformer encoder

2.2 The improved transformer encoder

Typically, the transformer consists of two parts: the encoder and the decoder. In this paper, we use an improved transformer encoder for feature fusion, as shown in Fig. 1.

According to Chen et al. [12], the positional encoding in the traditional transformer is not applicable in dual-path networks, so our transformer replaces one of the linear layers with a gated recurrent unit (GRU) layer based on the positional-wise feed-forward network, thereby better learning the positional information. The resulting feed-forward network (FFN), compared to traditional linear layers, incorporates nonlinear activation functions and additional layers, making it more flexible and capable of learning complex mappings. The method is shown in Eqs. 4 and 5:

$$\begin{aligned}\text{Res} &= \text{FFN}(\text{Mid}) \\ &= \text{Linear}(\text{ReLU}(\text{GRU}(\text{Mid})))\end{aligned} \tag{4}$$

$$\text{Output} = \text{LayerNorm}(\text{Mid} + \text{Res}) \tag{5}$$

where $\text{Mid} \in \mathbb{R}^{l \times d}$ is the input sequences with length l and dimension d, and $\text{FFN}(\cdot)$ denotes the output of the position-wise feed-forward network which contains $d \times 4$ hidden layer sizes.

2.3 The improved DenseBlock

The dilated DenseBlock, a core component of DenseNet [13], is initially applied in the field of computer vision and is renowned for its densely connected layers. This design ensures comprehensive feature propagation, which is crucial for maintaining detailed neighborhood feature integration across various image processing tasks.

Due to its ability to maintain a consistent gradient flow, the DenseBlock has also shown effective performance in audio processing tasks within the time–frequency domain. The dilated DenseBlock, incorporating convolution layers with varying dilation parameters, excels at integrating time and frequency domain features, which ensures the preservation of critical time–frequency information and enhances speech clarity.

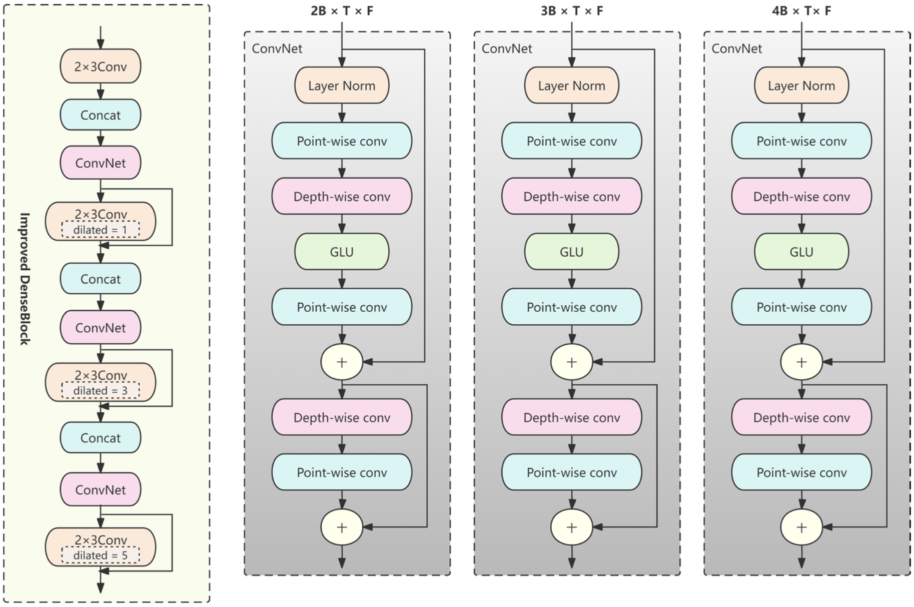


Fig. 2 The improved DenseBlock

In this paper, we optimize the DenseBlock using depth-wise convolution layers and point-wise convolution layers on the basis of dilated DenseBlock. This optimized module is implemented in both the encoder and the decoder of the Unet architecture for feature integration. The specific structure of this module is shown in Fig. 2.

3 Proposed methods

3.1 The time–frequency transformer

The transformer, acting as a critical component for feature fusion, has been extensively validated for its effectiveness in audio denoising for speech enhancement. This module is utilized within the bottleneck between the encoder and the decoder, as illustrated in Eq. 6, where the module takes the output of the encoder as input and provides denoised features as output to the decoder for spectrum reconstruction.

$$D_0 = \mathrm{TFT}(E_n) \tag{6}$$

In Eq. 6, TFT() represents the time–frequency transformer proposed in this paper. D_0 stands for the input of the decoder, and E_n denotes the output of the encoder.

Similar to the dual-path transformer proposed by Dang et al. [6], which alternates between full-band and sub-band transformers, we introduce the time–frequency transformer within the bottleneck of the Unet architecture. Following the full-band and sub-band transformers, a global transformer is added to further integrate features. Additionally, the 1×1 convolution layers are employed between the transformers to learn more two-dimensional features. The structure of the time–frequency transformer is depicted in (Fig. 3).

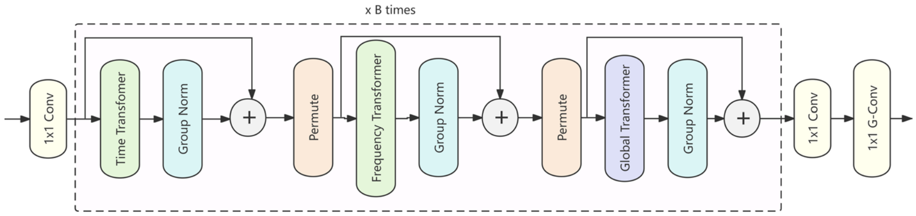


Fig. 3 The time–frequency transformer

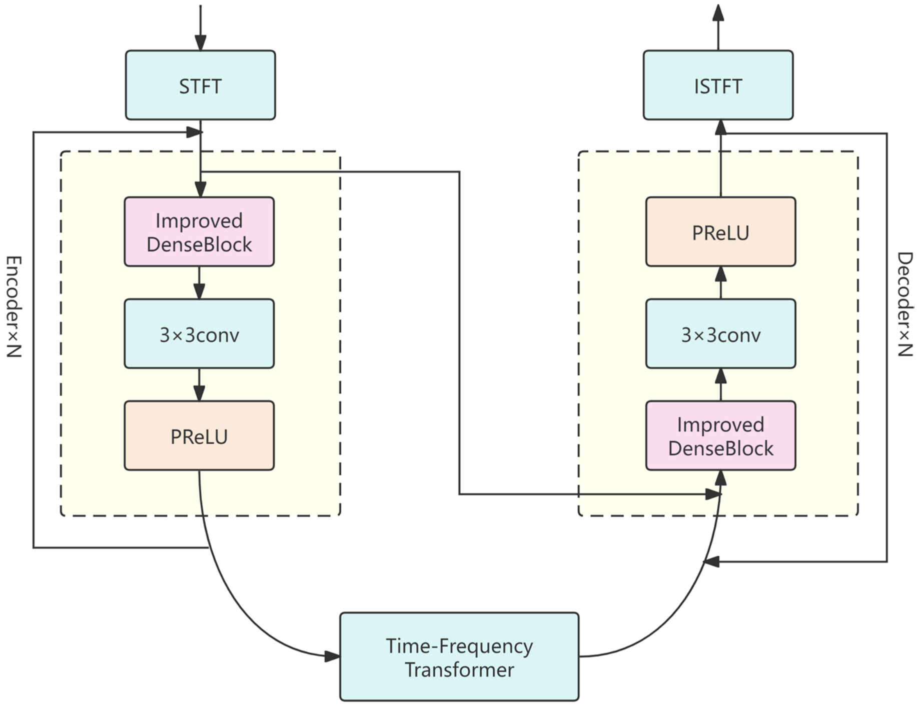


Fig. 4 The structure of TFDense-Net

3.2 TFDense-Net

This paper proposes TFDense-Net as the backbone structure for the enhancement network, based on the time–frequency transformer and the improved DenseBlock. TFDense-Net adopts a Unet architecture, primarily composed of three modules: the encoder, the decoder and the bottleneck, and it is depicted in Fig. 4.

The encoder employs the DenseBlock as the feature extraction module and utilizes 3×3 convolution layers for downsampling, where each downsampling operation reduces both the number of time frames and the number of frequency bins by half. The formulation is provided in Eq. 7:

$$\begin{aligned} E_i &= \text{Encoder}_i(E_{i-1}) \\ &= \text{Convolution}_i(IDB(E_{i-1})) \end{aligned} \tag{7}$$

where E_i is the output of the ith encoder. $IDB()$ is the improved DenseBlock layer, and $E_0 \in \mathbb{R}^{C \times T \times F}$ is the spectrum of the speech signal after STFT transformation. E_i

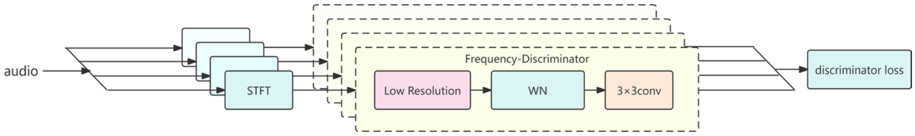


Fig. 5 The multi-spectrogram discriminator

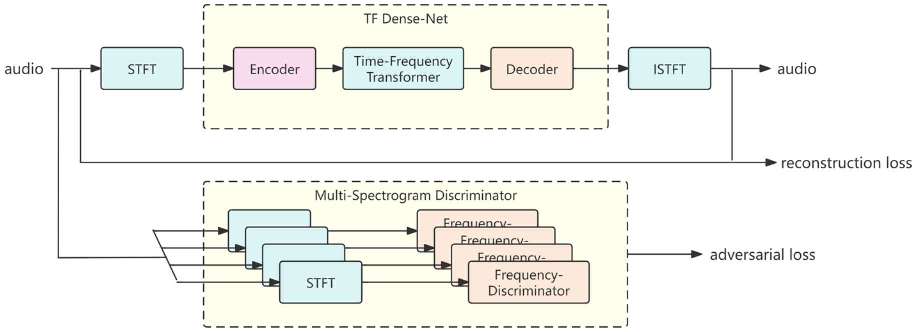


Fig. 6 The structure of TFDense-GAN

contains half of dimensions of time and frequency sizes of E_{i-1} after each downsampling operation.

In the decoder, upsampling is performed using transposed convolution with the same hyperparameters. Prior to upsampling, the improved DenseBlock is employed for feature extraction. After each upsampling operation, both the time and frequency domain lengths are doubled. The formulation is provided in Eq. 8:

$$\begin{aligned} D_i &= \text{Decoder}_i(D_{i-1}, E_i) \\ &= \text{TransposedConv}_i(IDB([D_{i-1}, E_i])) \end{aligned} \tag{8}$$

where D_i is the output of the ith decoder. $D_0 \in \mathbb{R}^{C\times T/2^n \times F/2^n}$ is the output of the bottleneck module, and n is the number of the encoder and decoder layers. After each upsampling operation, which includes an improved DenseBlock and a transposed convolutional layer, D_i contains twice the number of time frames and twice the number of frequency bins as E_{i-1}.

Based on the characteristics of Unet, the bottleneck can better aggregate speech features, and due to its usage after downsampling, the increase in model parameters and time complexity caused by the transformer is significantly reduced. This module can be trained independently, as well as jointly with the generator and discriminator, to achieve better results.

3.3 The multi-spectrogram discriminator

Building upon TFDense-Net, this paper introduces the multi-spectrogram discriminator as an enhanced discriminator, utilizing scoring-assisted evaluation as a metric for audio quality assessment. This metric not only evaluates the quality of the audio but also facilitates backpropagation to optimize training outcomes.

We apply STFT to speech signals under different window sizes and hop lengths. For each transformation result, we utilize a convolutional neural network (CNN) for further feature fusion and progressively downsample it to $\frac{1}{16}$ of its original size, yielding discriminator scores. The discriminator's loss function will be introduced in Sect. 3.5, and its architecture is depicted in Fig. 5.

3.4 TFDense-GAN

We present the overall structure of TFDense-GAN in Fig. 6. Initially, we apply STFT to the input speech signals with a fixed window size and hop length, followed by feeding them into TFDense-Net to compute the reconstruction loss. Building upon this framework, we optimize audio quality by comparing clean speech signals with denoised speech signals using our proposed multi-spectrogram discriminator. This method focuses on restoring fine details in sound quality that traditional loss functions may not adequately address. Comprising multiple STFTs with varying window sizes and hop lengths as inputs, the multi-spectrogram discriminator evaluates audio quality and feeds the information into the adversarial loss function.

3.5 The loss function

The loss function we employ comprises two components: the time domain loss and the time–frequency domain loss. This composite loss function enhances model performance from both waveform and spectrogram perspectives. It is defined as follows:

$$L_{\mathrm{TF}} = \alpha \times L_{\mathrm{waveform}} + \beta \times L_{\mathrm{spectrum}} \tag{9}$$

where α and β denote the weights assigned to the time domain loss and the time–frequency domain loss, respectively. L_{waveform} and L_{spectrum} represent the mean squared error of the audio in the time domain and that in the time–frequency domain, respectively. Their formulations are as follows:

$$L_{\mathrm{waveform}} = \mathbb{E}_{y,\hat{y}}[(y-\hat{y})^2] \tag{10}$$

$$L_{\mathrm{spectrum}} \quad = \mathbb{E}_{Y_{\mathrm{real}},\hat{Y}_{\mathrm{real}}}\left[(Y_{\mathrm{real}} - \hat{Y}_{\mathrm{real}})^2\right] + \mathbb{E}_{Y_{\mathrm{imag}},\hat{Y}_{\mathrm{imag}}}\left[(Y_{\mathrm{imag}} - \hat{Y}_{\mathrm{imag}})^2\right] \tag{11}$$

where y represents the clean speech, and $\hat{y}$ represents the denoised speech generated by our proposed model. Y and $\hat{Y}$ are the spectra obtained by applying STFT to y and $\hat{y}$, respectively, while Y_{real} and Y_{imag} represent the real and imaginary parts of Y, respectively.

Additionally, in GANs, adversarial training involves a minimization optimization task for the discriminator loss L_D and the corresponding generator loss L_{GAN}. These loss functions can be expressed as follows:

$$L_{GAN_{w,h}} = \mathbb{E}_{y,\hat{y}}\left[(D(y,\hat{y}) - 1)^2\right] \tag{12}$$

$$L_{\mathrm{GAN}} = \frac{1}{|S|} \times \sum_{(w,h)\in S} L_{\mathrm{GAN}_{w,h}} \tag{13}$$

$$L_{D_{w,h}} = \mathbb{E}_{y,\hat{y}}\left[(D(y,y)-1)^2 + (D(y,\hat{y}) - Q'(y,\hat{y}))^2\right] \tag{14}$$

$$L_D = \frac{1}{|S|} \times \sum_{(w,h)\in S} L_{D_{w,h}} \tag{15}$$

where D stands for the discriminator, and Q' indicates the normalized PESQ score, which maps the original range of −0.5 to 4.5 to a new range of 0–1. $L_{D_{w,h}}$ and $L_{\mathrm{GAN}_{w,h}}$ represent the two adversarial loss functions after STFT with window length w and hop length h, and S denotes the set of chosen window lengths and hop lengths.

The final generator loss is expressed as follows:

$$L_G = \gamma_1 L_{\mathrm{TF}} + \gamma_2 L_{\mathrm{GAN}} \tag{16}$$

where γ_1 and γ_2 represent the weights assigned to the corresponding loss functions.

4 Experimental setup

4.1 Datasets

In this paper, we utilize two public datasets that are widely employed for evaluating the performance of models in speech enhancement tasks.

VCTK + DEMAND. This dataset is one of the most frequently used datasets for speech enhancement. The clean audio subset comprises utterances from the VoiceBank corpus, which was originally proposed by Wang et al. [14]. The training subset includes 11,572 utterances from 28 speakers, while the test subset contains 872 utterances from 2 speakers. The noise set is extensive. Its training subset encompasses 40 different noise conditions featuring 10 types of noises-8 sourced from DEMAND as proposed by Joachim et al. [15], and 2 that are synthetically created-across SNRs of 0, 5, 10, and

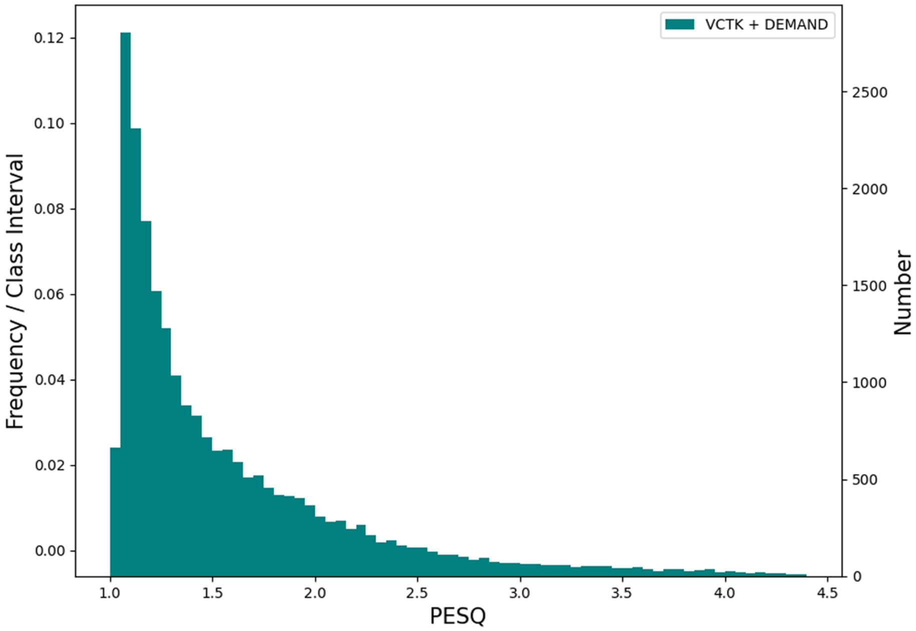


Fig. 7 The PESQ values of all audio samples and their proportions in the VCTK + DEMAND dataset

15 dB, and its test subset includes 20 different noise conditions with 5 types of unseen noises from DEMAND, presented at SNRs of 2.5, 7.5, 12.5, and 17.5 dB. To visually present the composition of the dataset, we have compiled statistics on the PESQ values of all audio samples in the training data and their proportions in the dataset. The results are shown in Fig. 7.

The Interspeech Deep Noise Suppression Challenge This dataset, proposed by Reddy et al. [16], is specifically designed for the Interspeech Deep Noise Suppression Challenge. It includes over 500 h of clean speech clips from 2150 speakers and more than 180 h of noise clips. For model evaluation, it features a non-blind validation set divided into two categories: with and without reverberation, each containing 150 noisy-clean pairs. Following the scripts provided by the organizers, we generate 3000 h of training data with SNRs ranging randomly from − 3 to 15 dB. To visually present the composition of the dataset, we have compiled statistics on the PESQ values of all audio samples in the training data and their proportions within the dataset. The results are shown in Fig. 8.

4.2 Training setup

For all training audio data, we initially resample them at 16,000 sampling rate and standardize the length of each audio signal to 2 s, resulting in a total of 32,000 sampling points per signal. If the length of the original audio signal is less than 2 s, we pad the signal with zeros at the end. Conversely, if the signal length exceeds 2 s, we randomly crop a 2-s segment as the training sample for the current epoch, ensuring a uniform signal length within the same batch. Additionally, we filter out audio data with low information entropy to enhance the purity of the training data.

We use the Adam optimizer [17] with $\beta_1 = 0.9$, $\beta_2 = 0.99$ and apply gradient clipping with a maximum L2-norm of 5 to avoid gradient explosion. During the training stage, we

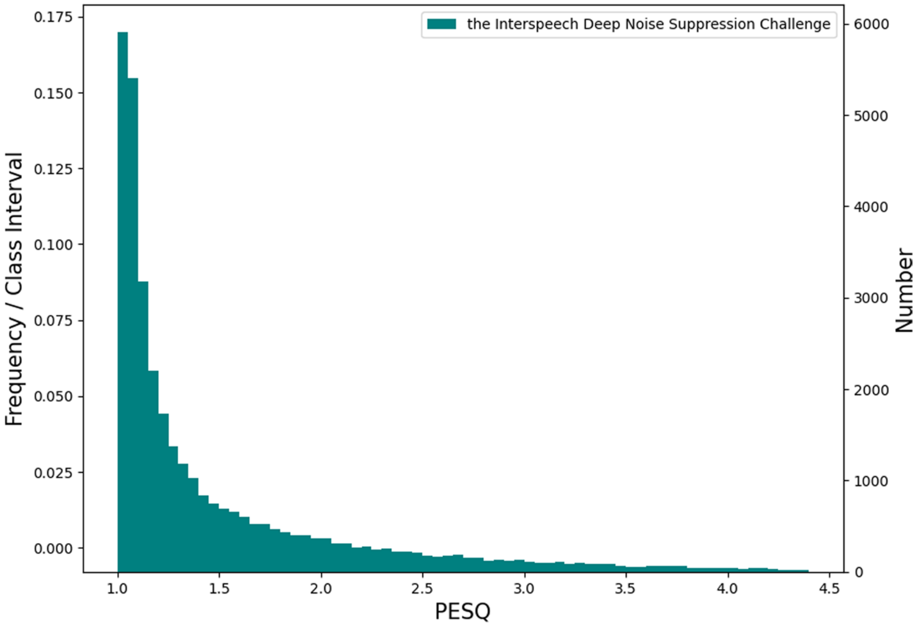


Fig. 8 The PESQ values of all audio samples and their proportions in the Interspeech Deep Noise Suppression Challenge dataset

train all of our proposed models for 300 epochs, with each epoch consisting of 20,000 speech samples. Each sample is randomly cropped into 1-s segments during the dataset loading process. The window length and FFT size for STFT and iSTFT are set to 512, with a frame shift of 256.

The number of feature maps C of the time–frequency (T-F) spectrum is set to 64. The depth of both the encoder and the decoder is set to 3. Each improved DenseBlock layer comprises one dilated convolution layer with a dilated rate of $d = 2$, a deep-wised convolution layer, and a point-wised convolution layer for feature fusion. The number of input channels in the successive layers of the DenseBlock increases linearly as C, $2C$, $3C$, and $4C$, with the output after each convolution layer consisting of C channels. We use 4 deep dual-path transformers, indicating $B = 4$ and $h = 4$. During the training phase, we train the proposed models for 200 epochs. Experiments are conducted using one NVIDIA GeForce RTX 3090 GPU, and the models are validated with the publicly available datasets mentioned in Sect. 4.1. The specific experimental hardware and software configuration is detailed in Table 1.

To adjust the learning rate during the training phase, we adopt a dynamic strategy proposed by Vaswani et al. [10], with modifications tailored to our training process. The formula is as follows:

$$lr = \begin{cases} k_1 * d_{\text{model}}^{-0.5} * n * warmup^{-1.5} & n \leq warmup \\ k_2 * 0.98^{\text{epoch}/2} & n > warmup \end{cases} \tag{17}$$

where n is the number of steps, d_{model} denotes the feature size of the input of the transformer, and k_1 and k_2 are tunable scalars. In our experiments, we set $d_{\text{model}} = 64$, $k_1 = 0.2, k_2 = 4e^{-4}$ and $warmup = 4,000$.

4.3 Evaluation metrics

In this paper, we use STOI, PESQ, Si-SDR, CBAK, COVL, CSIG, FLOPs, and Causal as evaluation metrics for comparison.

STOI, which is the abbreviation of short-time objective intelligibility, is a metric used to assess the intelligibility of speech signals, with values ranging from 0 to 1 [18]. It was developed to provide a robust method for evaluating the clarity and understandability of speech signals, especially in conditions where there is noise or other types of distortions.

PESQ, which stands for perceptual evaluation of speech quality, is a standard used metric for assessing the quality of speech signals, often in telecommunications [19]. It aims at providing an objective measure that can correspond well to subjective

Table 1 Hardware and software configuration

Configuration	Detail
CPU	AMD Ryzen 5 5600G with Radeon graphics
RAM	32G
Graphics card	NVIDIA GeForce RTX 3090
Operating system	64-bits Ubuntu 22.04.1
CUDA	12.4
Programming language	Python 3.11

evaluations of speech quality. PESQ is designed to work in a wide variety of network conditions and for many types of distortions including noise, codec distortions, and other network impairments. Its values range from -0.5 to 4.5.

Si-SDR is a metric used to evaluate the performance of audio source separation algorithms [20]. Source separation is the task of isolating individual audio sources from a mixture of sounds. Si-SDR provides a way to quantify how well a separation algorithm has performed, by comparing the separated signals to the original, ground-truth signals. The formulas for Si-SDR are as follows:

$$s_{\text{target}} = \frac{< \hat{s}, s > s}{||s||^2} \tag{18}$$

$$e_{\text{noise}} = \hat{s} - s_{\text{target}} \tag{19}$$

$$\text{SiSDR} = 10 \log_{10} \frac{||s_{\text{target}}||^2}{||e_{\text{noise}}||^2} \tag{20}$$

where $< \hat{s}, s >$ represents a scalar, which is the dot product of $< \hat{s}, s >$. $||s||$ is the L2-norm of s.

CBAK refers to the mean opinion score (MOS) prediction of the intrusiveness of background noise within a speech signal. It serves as an index for evaluating the efficacy of background noise suppression in speech signals, with a scale ranging from 0 to 5, where higher values closer to 5 indicate superior noise suppression. This metric predominantly focuses on the relationship between the clarity and comprehensibility of the speech signal and the presence of background noise.

COVL, representing the MOS prediction of the overall effect, measures the degree of distortion caused by overloading within a speech signal [21]. Overload distortion can result in speech deformation or loss of information, adversely affecting the clarity and comprehensibility of the speech.

CSIG, the MOS predictor of speech distortion, assesses the quality of the speech signal itself, excluding the influence of background noise and other distortive elements. This metric is primarily concerned with the clarity and naturalness of the speech.

FLOPs, which stands for Floating Point Operations per Second, is a metric that quantifies the number of floating-point arithmetic operations executed by a system or model within one second. This measure is widely employed to assess the computational complexity of models and algorithms, offering valuable insights into the computational resources necessary for executing tasks such as inference and training in machine learning, as well as signal processing.

Causal refers to the property of a system or model in which the output at any given time depends solely on the current and past inputs, rather than on any future inputs. In speech processing, *causal* models are essential for real-time applications, where future speech information is not available at the moment of processing.

Table 2 The experimental results with different values of *d* and *l* on the VCTK + DEMAND dataset

Model	d	l	PESQ	STOI	Si-SDR	CBAK	COVL	CSIG
TFDense-Net	32	4	3.35	0.95	19.24	3.27	4.01	4.35
TFDense-Net	32	3	3.32	0.95	18.57	3.19	4.00	4.33
TFDense-Net	128	2	3.39	0.96	19.01	3.41	4.09	4.41
TFDense-Net	128	3	3.43	0.96	20.12	3.47	4.19	4.43
TFDense-Net	64	2	3.33	0.96	17.67	3.39	4.10	4.32
TFDense-Net	64	3	3.40	0.96	19.57	3.47	4.13	4.37
TFDense-Net	64	4	3.42	0.96	**20.13**	3.51	4.17	4.50
TFDense-GAN	32	4	3.41	0.95	15.35	3.41	4.16	4.52
TFDense-GAN	32	3	3.38	0.94	14.57	3.37	4.13	4.48
TFDense-GAN	128	2	3.46	0.97	16.13	3.69	4.22	4.55
TFDense-GAN	128	3	3.59	0.97	17.28	3.73	4.27	4.72
TFDense-GAN	64	2	3.56	0.97	18.39	3.64	4.26	4.69
TFDense-GAN	64	3	3.60	0.97	18.77	3.79	**4.34**	4.75
TFDense-GAN	64	4	**3.62**	**0.97**	18.96	**3.86**	4.33	**4.80**

The best results are shown in bold

Table 3 The experimental results with different values of *d* and *l* on the Interspeech Deep Noise Suppression Challenge dataset

Model	d	l	PESQ	STOI	Si-SDR	CBAK	COVL	CSIG
TFDense-Net	32	4	3.39	0.95	20.23	3.27	4.02	4.39
TFDense-Net	32	3	3.34	0.96	19.37	3.22	3.96	4.31
TFDense-Net	128	2	3.41	0.96	19.41	3.42	4.10	4.25
TFDense-Net	128	3	3.49	0.96	20.43	3.51	4.19	4.45
TFDense-Net	64	2	3.35	0.96	18.33	3.41	4.07	4.39
TFDense-Net	64	3	3.41	0.96	19.91	3.49	4.20	4.38
TFDense-Net	64	4	3.47	0.96	**20.91**	3.55	4.21	4.47
TFDense-GAN	32	4	3.40	0.95	17.33	3.44	4.17	4.56
TFDense-GAN	32	3	3.41	0.95	15.17	3.33	4.23	4.51
TFDense-GAN	128	2	3.50	0.96	18.23	3.69	4.25	4.52
TFDense-GAN	128	3	3.62	0.97	19.63	3.71	4.30	4.75
TFDense-GAN	64	2	3.58	0.97	18.77	3.77	4.36	4.72
TFDense-GAN	64	3	3.62	0.97	20.37	3.85	4.33	4.81
TFDense-GAN	64	4	**3.66**	**0.98**	20.41	**3.90**	**4.37**	**4.84**

The best results are shown in bold

5 Experimental results

In this section, we conduct experiments on the VCTK + DEMAND dataset and the Interspeech Deep Noise Suppression Challenge dataset, respectively, to evaluate the performance of our models. First, we perform hyperparameter analysis to select suitable hyperparameters for TFDense-Net and TFDense-GAN. Next, we conduct ablation studies to demonstrate the effectiveness of our proposed time–frequency transformer and to show the noise reduction improvements achieved by TFDense-GAN, which incorporates the multi-spectrogram discriminator on top of TFDense-Net. Additionally, we compare TFDense-GAN with other state-of-the-art models on the two aforementioned datasets. Lastly, to visually display the denoising effectiveness of TFDense-GAN, we present and analyze several spectrograms.

5.1 Hyperparameter analysis

To assess the impact of the input dense dimension d and the number of dense layers l in both the encoder and decoder on the performance of TFDense-Net and TFDense-GAN, we designed seven experiments with varying values of d and l for both models. These experiments were conducted on the VCTK + DEMAND dataset and the Interspeech Deep Noise Suppression Challenge dataset. The results are presented in Tables 2 and 3, respectively.

According to Tables 2 and 3, we observe a significant decrease in model performance when $d = 32$ or $l = 2$, which is likely due to insufficient feature extraction. Additionally, when $d = 128$, the performance does not improve, which may be attributed to overfitting on the datasets. However, smaller hyperparameters result in fewer model parameters and faster convergence. Therefore, we find it more optimal to set $d = 64$ and $l = 4$.

In our experiments, we also observe that the Si-SDR scores are lower when using the multi-spectrogram discriminator (i.e., TFDense-GAN) compared to when it is not used (i.e., TFDense-Net). This discrepancy is related to the requirement in the Si-SDR formula for the generated speech and target speech timestamps to be aligned. The loss function employed in GANs can lead to frame-level content shifts in the generated speech, which might result in better perceived audio quality but relatively lower Si-SDR scores.

5.2 Ablation experiments

To assess the efficacy of the proposed TFDense-Net and the time–frequency transformer, as well as the impact of the multi-spectrogram discriminator on model performance, we conducted ablation experiments using the VCTK + DEMAND dataset and the Interspeech Deep Noise Suppression Challenge dataset. The results of these experiments are presented in Tables 4 and 5, respectively.

There are four groups of experiments presented in Tables 4 and 5. In the "TFDense-Net" group, we validate the denoising performance of TFDense-Net without the involvement of the multi-spectrogram discriminator. The experiment not only demonstrates the favorable denoising performance of our proposed TFDense-Net but also compares it with the model using the multi-spectrogram discriminator,

Table 4 The ablation experimental results on the VCTK + DEMAND dataset

Model	PESQ	STOI	Si-SDR	CBAK	COVL	CSIG
TFDense-Net	3.42	0.96	**20.13**	3.51	4.17	4.50
TFDense-Net-dual-path	3.35	0.96	19.36	3.38	4.01	4.31
TFDense-Net-dual-path+MSD	3.55	0.96	18.58	3.63	4.19	4.62
TFDense-GAN	**3.62**	**0.97**	18.96	**3.86**	4.33	**4.80**

The best results are shown in bold

Table 5 The ablation experimental results on the Interspeech Deep Noise Suppression Challenge dataset

Model	PESQ	STOI	Si-SDR	CBAK	COVL	CSIG
SNR = 15.0						
TFDense-Net	4.07	0.99	21.55	3.92	4.27	4.89
TFDense-Net-dual-path	4.05	0.98	20.18	3.81	4.23	4.86
TFDense-Net-dual-path+MSD	3.99	0.98	24.35	3.58	4.09	4.51
TFDense-GAN	4.02	0.98	25.67	3.73	4.20	4.72
SNR = 10.0						
TFDense-Net	3.73	0.97	23.77	3.56	4.11	4.42
TFDense-Net-dual-path	3.68	0.97	22.95	3.49	4.02	4.41
TFDense-Net-dual-path+MSD	3.82	0.98	19.98	3.74	4.16	4.47
TFDense-GAN	3.85	0.98	20.75	3.79	4.10	4.50
SNR = 5.0						
TFDense-Net	3.57	0.97	20.63	3.63	4.02	4.52
TFDense-Net-dual-path	3.55	0.96	20.17	3.58	3.98	4.33
TFDense-Net-dual-path+MSD	3.60	0.97	19.12	3.81	4.12	4.56
TFDense-GAN	3.62	0.97	19.15	3.92	4.20	4.53
SNR = 0						
TFDense-Net	3.01	0.95	17.59	2.95	3.18	3.38
TFDense-Net-dual-path	2.93	0.94	17.39	2.88	3.09	3.36
TFDense-Net-dual-path+MSD	3.05	0.95	15.55	2.91	3.12	3.40
TFDense-GAN	3.08	0.96	16.37	2.99	3.15	3.49
SNR = -5.0						
TFDense-Net	2.41	0.94	15.25	2.39	2.87	2.98
TFDense-Net-dual-path	2.39	0.93	14.93	2.20	2.68	2.93
TFDense-Net-dual-path+MSD	2.55	0.93	11.77	2.51	2.91	3.03
TFDense-GAN	2.63	0.95	12.08	2.53	2.93	3.05

specifically TFDense-GAN, highlighting an enhancement in denoising performance when the multi-spectrogram discriminator is employed. In the "TFDense-Net-dual-path" group, we replace the proposed time–frequency transformer with the dual-path transformer. In the "TFDense-Net-dual-path+MSD" group, we augment TFDense-Net-dual-path with the multi-spectrogram discriminator (MSD) to demonstrate the general applicability of the multi-spectrogram discriminator. Lastly, we conduct experiments using our proposed "TFDense-GAN," which integrates TFDense-Net and the multi-spectrogram discriminator. By comparing TFDense-GAN with the previous three groups of experiments, we aim to verify the impact of both the time–frequency transformer and the multi-spectrogram discriminator on model performance.

From Table 4, we observe that TFDense-Net exhibits promising denoising performance in a non-self-supervised training environment. However, when comparing the experimental groups of TFDense-Net and TFDense-GAN, we find that the introduction of the multi-spectrogram discriminator significantly enhances

Table 6 Comparison with other state-of-the-art models on the VCTK + DEMAND dataset

Model	**Year**	**Param.**	**PESQ**	**STOI (%)**	**Si-SDR (dB)**	**FLOPs**	**Causal**
Noisy	**–**	**n/a**	**1.97**	**0.91**	**–**	**–**	**–**
PHASEN [28]	2020	20.9M	2.99	–	10.08	206G	No
DCCRN [25]	2020	3.67M	2.57	0.94	19.13	**25.2G**	Yes
DEMUCS [22]	2020	128M	3.07	0.95	–	77.8G	–
MetricGAN+ [32]	2021	2.6M	3.15	0.93	–	–	–
TSTNN [23]	2021	0.92M	2.96	0.95	18.82	–	–
SE-Conformer [24]	2021	–	3.13	0.95	–	–	–
DB-AIAT [26]	2022	2.81M	3.31	0.96	10.79	68G	–
DPT-FSNet [6]	2022	**0.88M**	3.33	0.96	18.85	55.7G	–
CMGAN [31]	2022	1.83M	3.41	0.96	11.10	116G	–
CleanUnet [4]	2022	46.07M	2.905	0.956	–	–	Yes
D2Net [27]	2022	1.13M	3.27	0.96	19.78	–	–
TridenSE [29]	2022	3.03M	3.47	0.96	–	59.8G	No
MP-SENet [30]	2023	2.26M	3.6	0.96	–	–	–
TFDense-Net	2024	3.63M	3.42	0.96	**20.13**	61.2G	Yes
TFDense-GAN	2024	3.63M	**3.62**	**0.97**	18.96	63.7G	Yes

The best results are shown in bold

the performance of speech enhancement tasks. Additionally, compared to the dual-path transformer, our proposed time–frequency transformer achieves more effective feature fusion, resulting in superior performance.

In Table 5, the Interspeech Deep Noise Suppression Challenge dataset is used to generate training and test sets with different SNRs. We evaluate the denoising performance on datasets with varying SNRs. We find that in high-SNR environments, the results of the four experimental groups are comparable. Notably, in the SNR = 15 dB experiment, the performance of TFDense-Net is even better than that of TFDense-GAN. However, under low-SNR conditions, TFDense-GAN exhibits significant advantages. Additionally, the performance of our proposed time–frequency transformer surpasses that of the dual-path transformer, underscoring its effectiveness. Nevertheless, our models show noticeable improvements at SNR = 5 dB and above, whereas the evaluation metrics such as PESQ, STOI and Si-SDR perform poorly in the SNR = 0 dB and SNR = -5 dB experiments, indicating that there is still room for improvements in our models.

5.3 Comparison with other state-of-the-art models

5.3.1 Experimental results on the VCTK + DEMAND dataset

We objectively compare TFDense-Net and TFDense-GAN with other state-of-the-art denoising models on the VCTK + DEMAND dataset, as shown in Table 6. For time domain methods, we include standard models such as CleanUnet [4], DEMUCS [22], TSTNN [23], and SE-Conformer [24]. TFDense-GAN demonstrates a significant reduction in parameters while achieving notable performance improvements. When compared with time–frequency domain supervised learning models like DPT-FSNet [6], DCCRN [25], DB-AIAT [26], and D2Net [27], TFDense-GAN achieves improvements across all three evaluation metrics while maintaining a comparable number of parameters. Additionally, TFDense-GAN significantly outperforms PHASEN [28] in

Table 7 Comparison with other state-of-the-art models on the Interspeech Deep Noise Suppression Challenge dataset

Model	Year	Param.	PESQ	STOI (%)	Si-SDR (dB)	FLOPs	Causal
Noisy	–	n/a	1.97	91.5	–	–	–
NSNet [16]	2020	–	2.15	94.47	15.61	–	–
DTLN [33]	2020	–	–	94.76	16.34	–	Yes
PoCoNet [34]	2020	-	2.75	–	–	–	No
FullSubNet [35]	2021	5.6M	2.77	96.11	17.29	–	–
CTS-Net [36]	2021	4.99M	2.94	96.66	17.99	–	Yes
DPT-FSNet [6]	2022	**0.88M**	3.26	97.68	20.36	51.5G	–
CleanUNet [4]	2022	46.07M	3.146	95.6	–	–	Yes
FRCRN [37]	2022	6.9M	3.27	97.69	19.78	–	Yes
GaGNet [38]	2022	5.94M	3.17	97.13	18.91	**8.13G**	Yes
TaylorSENet [39]	2022	5.4M	3.22	97.36	19.15	–	Yes
MFNet [40]	2023	-	3.43	97.98	20.31	–	No
MP-SENet [30]	2023	2.26M	3.62	98.16	21.03	–	–
TFDense-Net	2024	3.63M	3.47	96.47	**20.91**	58.4G	Yes
TFDense-GAN	2024	3.63M	**3.66**	**98.25**	20.41	59.8G	Yes

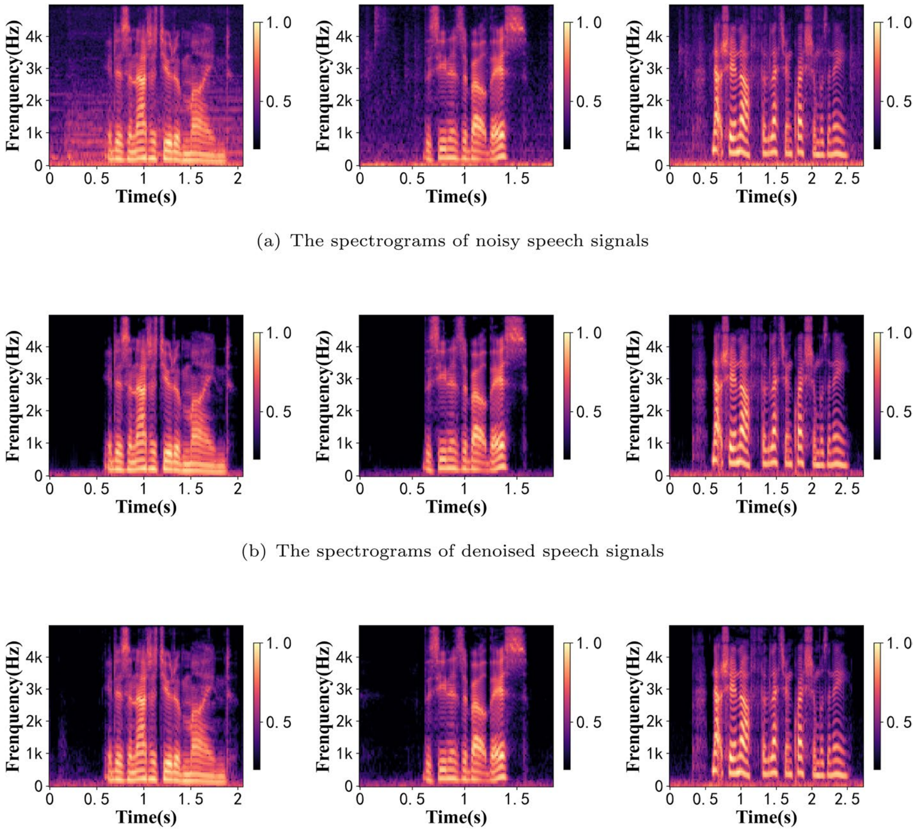
The best results are shown in bold

terms of parameters and surpasses TridenSE [29] and MP-SENet [30] in terms of STOI. Furthermore, although TFDense-GAN is a causal model, it achieves superior denoising performance compared to non-causal models such as PHASEN [28] and TridenSE [29]. In addition, TFDense-GAN demonstrates significantly lower computational complexity (FLOPs) compared to models like CMGAN [31] and DEMUCS [22], highlighting its efficiency in real-time applications. Finally, compared with GAN-based networks like MetricGAN+ [32] and CMGAN [31], TFDense-GAN achieves superior denoising performance and audio quality, further validating its competitiveness for noise suppression tasks.

5.3.2 Experimental results on the Interspeech Deep Noise Suppression Challenge dataset

In Table 7, we provide the experimental results of TFDense-Net and TFDense-GAN on the Interspeech Deep Noise Suppression Challenge dataset compared with other state-of-the-art models. It is evident that our models exhibit excellent denoising performance compared to the unprocessed noisy speech signals (the first row “Noisy”). Compared with CleanUNet [4], our models have fewer parameters and achieve improvements across all evaluation metrics.

In terms of the PESQ metric, TFDense-GAN outperforms models such as FRCRN [37], GaGNet [38], and MFNet [40], while maintaining efficiency in parameters. Regarding the STOI metric, TFDense-GAN shows significant advancements over leading models such as DPT-FSNet [6], FullSubNet [35], and CTS-Net [36], highlighting its advantages in enhancing speech intelligibility. By further examining the Si-SDR metric, it is noted that the results may decline when using GAN (compared with MP-SENet [30]) due to the impact of frame shifts on the calculation. Nevertheless, thanks to the inclusion of the reconstruction loss, TFDense-GAN still achieves commendable performance in this metric. It shows substantial improvements over CNN-based models like PoCoNet [34]



(a) The spectrograms of noisy speech signals

(b) The spectrograms of denoised speech signals

(c) The spectrograms of clean speech signals

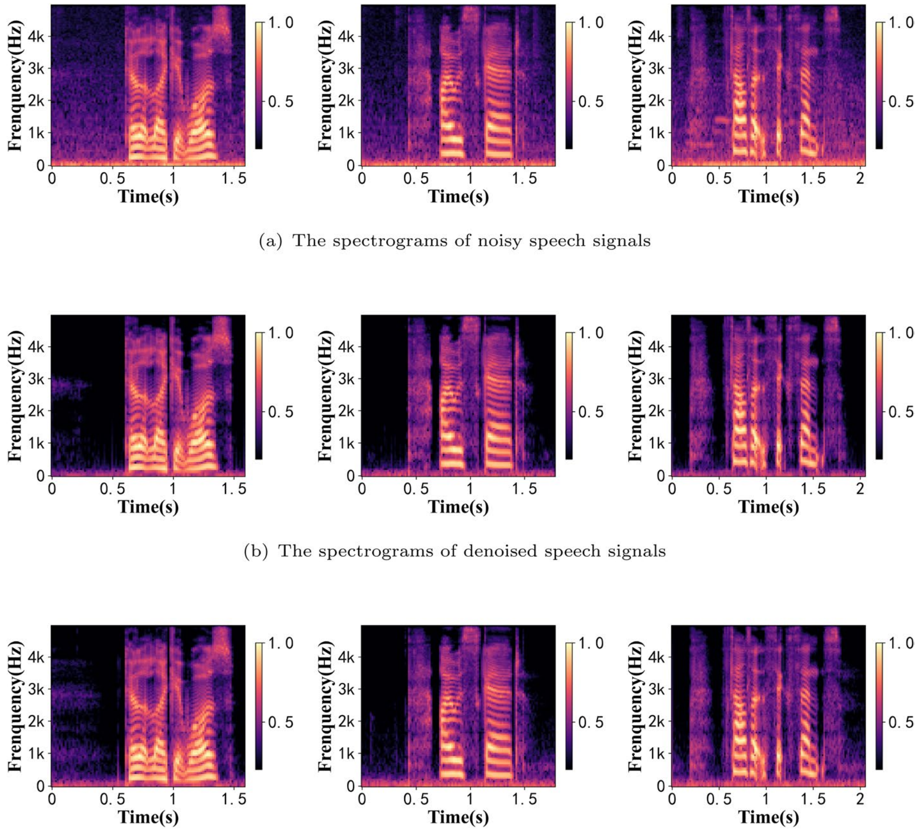
Fig. 9 Spectrograms sampled from the VCTK + DEMAND test set. The PESQ values of all the three noisy speech signals are greater than 2.0

and TaylorSENet [39] and also outperforms RNN-based models such as NSNet [16] and DTLN [33]. Although TFDense-GAN has 59.8G FLOPs, slightly higher than some causal models such as GaGNet [38], it achieves significantly superior performance in PESQ and STOI. At the same time, compared to large-parameter models like CleanUnet [4], it demonstrates lower computational complexity and better performance, showcasing an excellent balance between causal design and efficiency.

5.4 Spectrogram analysis

To provide a more intuitive evaluation of TFDense-GAN, we present spectrograms of denoised speech signals along with their corresponding noisy and clean speech signals in Figs. 9, 10, and 11. These speech signals are sampled from the VCTK + DEMAND test set and the Interspeech Deep Noise Suppression Challenge test set.

Specifically, in Fig. 9, we show spectrograms of speech signals with PESQ values exceeding 2.0, indicative of higher signal-to-noise ratios. Under low-noise conditions, our model demonstrates strong performance, with PESQ values for the denoised speech signals exceeding 4.0, reaching up to 4.5. In Fig. 10, the spectrograms depict noisy speech signals exhibiting PESQ values below 1.4, representing signals with low



(a) The spectrograms of noisy speech signals

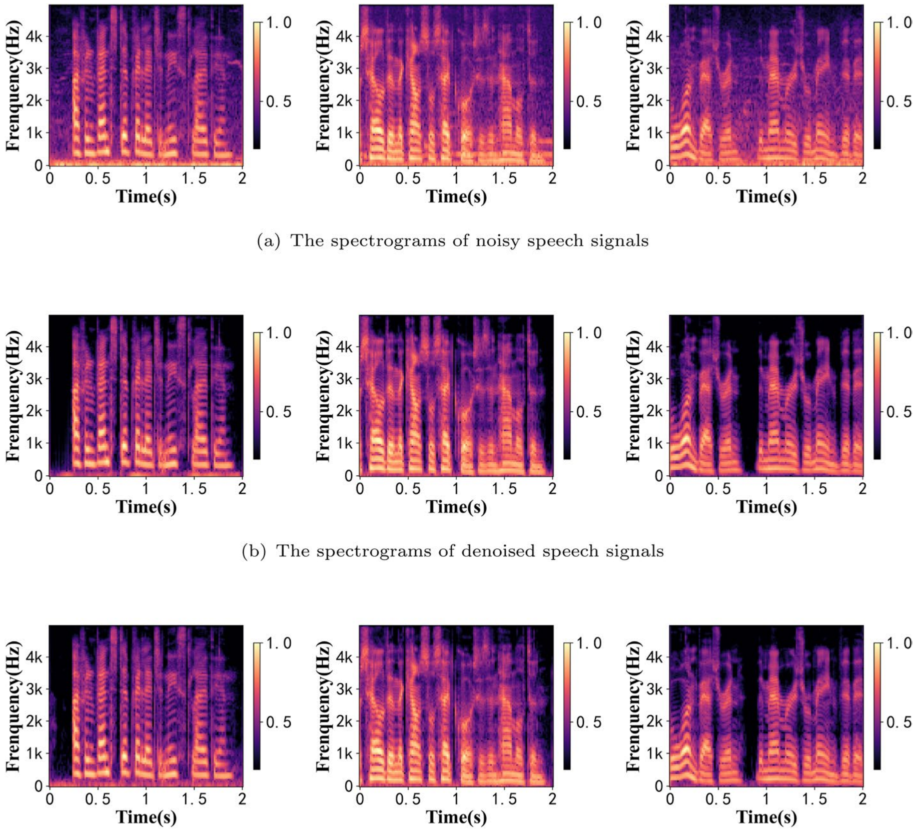
(b) The spectrograms of denoised speech signals

(c) The spectrograms of clean speech signals

Fig. 10 Spectrograms sampled from the VCTK + DEMAND test set. The PESQ values of all the three noisy speech signals are less than 1.4

signal-to-noise ratios. In high-noise environments, our model effectively restores audio intelligibility, enhancing the feasibility of downstream tasks. Nevertheless, some residual noise remains in the spectrograms. Overall, these spectrograms underscore TFDense-GAN's capability to filter out mid-frequency and low-frequency noise across different bands. Furthermore, by comparing the spectrograms of the clean speech signals, we observe that TFDense-GAN is able to rectify impurities in the speech signals, demonstrating its denoising capabilities beyond merely mapping based on references of the clean speech signals.

In Fig. 11, we present spectrograms of speech signals with PESQ values ranging from 1.4 to 2.0, indicative of moderate signal-to-noise ratios. Compared to high-SNR environments, the denoising performance under moderate SNR conditions shows a slight decline, as evidenced by a marginal decrease in PESQ values and the presence of some residual noise. Despite these differences, the model still effectively enhances the clarity of the speech signals and preserves more speech details during restoration. The denoising results in moderate SNR environments exhibit clear distinctions compared to low-SNR environments. In these moderate conditions, the model not only removes noise but also retains more detailed information in the speech signal,



(a) The spectrograms of noisy speech signals

(b) The spectrograms of denoised speech signals

(c) The spectrograms of clean speech signals

Fig. 11 Spectrograms sampled from the Interspeech Deep Noise Suppression Challenge test set. The PESQ values of all the three noisy speech signals are greater than 1.4 but less than 2.0

which can result in more natural and clear recovered speech, thereby enhancing users' understanding and perception of the speech content.

6 Limitations

During our experiments, we primarily tested the model on datasets consisting of speech signals mixed with noise. However, in real-world environments, speech signals are often affected not only by noise but also by factors such as reverberation, clipping, and downsampling. These factors can significantly alter speech structures and sampling rates, thereby posing challenges for model generalization.

Additionally, training GAN-based models like TFDense-GAN can sometimes suffer from instability, which may impact convergence and performance consistency. The computational complexity of the GAN framework, coupled with the increased number of parameters compared to baseline models, also introduces challenges in terms of resource requirements. This limitation affects the feasibility of deploying TFDense-GAN on devices with constrained computational capacities, such as mobile or embedded systems.

Moreover, the absence of real-time performance evaluations remains a notable limitation. Although our experiments have demonstrated strong results using offline metrics, such as PESQ and STOI, further work is necessary to evaluate and optimize the model's latency in real-time scenarios.

7 Conclusions and future work

In this paper, we have proposed TFDense-Net and its improved version, TFDense-GAN, for the tasks of speech enhancement. The experimental results on the VCTK + DEMAND dataset and the Interspeech Deep Noise Suppression Challenge dataset demonstrate that TFDense-GAN outperforms most of the current state-of-the-art models.

Future work will focus on addressing the limitations described in Sect. 6. First, we plan to incorporate reverberation and other distortions into the training data to enhance the model's robustness in complex acoustic conditions. Additionally, we will explore techniques to stabilize GAN training and reduce computational costs without compromising performance, as well as design lightweight versions of the architecture to facilitate deployment on resource-constrained devices. Furthermore, we will expand the scope of evaluations to include real-time scenarios and optimize the model for latency-sensitive applications. By addressing these limitations, we aim to enhance the practical applicability and generalizability of TFDense-GAN in diverse real-world conditions.

Acknowledgements
Not applicable.

Author contributions
Haoxiang Chen conceived the idea, implemented the model, and was the primary contributor to the paper. Yaogang Fu participated in the model implementation and conducted some experiments. Jinxiu Zhang and Xintong Zhou were responsible for collecting and analyzing the relevant datasets and contributed to the writing of the paper. Ruilong Wang, Yanyan Xu, and Dengfeng Ke supervised the project, providing critical revisions and guidance. All authors reviewed and approved the final manuscript.

Funding
This work was funded by the research project with Project No. 2023HXKFXX0024.

Availability of data and materials
All datasets used in this article are publicly available. The authors confirm that the data supporting the findings of this study are available within the article or its supplementary materials. The VCTK + DEMAND dataset is available at https://datashare.ed.ac.uk/handle/10283/2791. The Interspeech Deep Noise Suppression Challenge dataset is available at https://www.microsoft.com/en-us/research/academic-program/deep-noise-suppression-challenge-interspeech-2020/.

Declarations

Competing interests
The authors declare that they have no conflict of interest.

Received: 19 August 2024 Accepted: 17 February 2025
Published online: 22 March 2025

References

1. D. Rethage, J. Pons, X. Serra, A wavenet for speech denoising, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2018), pp. 5069–5073
2. A. Pandey, D. Wang, A new framework for cnn-based speech enhancement in the time domain. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(7), 1179–1188 (2019)
3. Y. Luo, N. Mesgarani, Conv-tasnet: surpassing ideal time-frequency magnitude masking for speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(8), 1256–1266 (2019)
4. Z. Kong, W. Ping, A. Dantrey, B. Catanzaro, Speech denoising in the waveform domain with self-attention, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 7867–7871

5. J. Kim, M. El-Khamy, J. Lee, T-gsa: transformer with gaussian-weighted self-attention for speech enhancement, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2020), pp. 6649–6653
6. F. Dang, H. Chen, P. Zhang, Dpt-fsnet: dual-path transformer based full-band and sub-band fusion network for speech enhancement, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 6857–6861
7. R. Ma, S. Li, B. Zhang, L. Fang, Z. Li, Flexible and generalized real photograph denoising exploiting dual meta attention. IEEE Trans. Cybern. **53**(10), 6395–6407 (2022)
8. S. Zhao, T.H. Nguyen, B. Ma, Monaural speech enhancement with complex convolutional block attention module and joint time frequency losses, in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2021), pp. 6648–6652
9. H. Liu, Q. Kong, Q. Tian, Y. Zhao, D. Wang, C. Huang, Y. Wang, Voicefixer: toward general speech restoration with neural vocoder. arXiv preprint arXiv:2109.13731 (2021)
10. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need. Adv. Neural Inf. Process. Syst. **30** (2017)
11. R. Ma, Y. Zhang, B. Zhang, L. Fang, D. Huang, L. Qi, Learning attention in the frequency domain for flexible real photograph denoising. IEEE Trans. Image Process. (2024)
12. J. Chen, Q. Mao, D. Liu, Dual-path transformer network: direct context-aware modeling for end-to-end monaural speech separation, in *Interspeech 2020*, pp. 2642–2646 (2020)
13. F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, K. Keutzer, Densenet: Implementing efficient convnet descriptor pyramids. arXiv preprint arXiv:1404.1869 (2014)
14. C. Veaux, J. Yamagishi, S. King, The voice bank corpus: design, collection and data analysis of a large regional accent speech database, in *2013 International Conference Oriental COCOSDA Held Jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*. IEEE (2013)
15. J. Thiemann, N. Ito, E. Vincent, The diverse environments multi-channel acoustic noise database: a database of multichannel environmental noise recordings. J. Acoust. Soc. Am. **133**(5), 3591 (2013)
16. C.K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, et al. The interspeech 2020 deep noise suppression challenge: datasets, subjective testing framework, and challenge results. arXiv preprint arXiv:2005.13981 (2020)
17. D.P. Kingma, J. Ba, Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
18. C.H. Taal, R.C. Hendriks, R. Heusdens, J. Jensen, An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Trans. Audio Speech Lang. Process. **19**(7), 2125–2136 (2011)
19. A.W. Rix, J.G. Beerends, M.P. Hollier, A.P. Hekstra, Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 749–752 (2001)
20. J.L. Roux, S. Wisdom, H. Erdogan, J.R. Hershey, Sdr–Half-baked or well done?, in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK*, pp. 626–630 (2019). https://doi.org/10.1109/ICASSP.2019.8683855
21. Y. Hu, P.C. Loizou, Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio Speech Lang. Process. **16**(1), 229–238 (2007)
22. A. Defossez, G. Synnaeve, Y. Adi, Real time speech enhancement in the waveform domain. arXiv preprint arXiv:2006.12847 (2020)
23. K. Wang, B. He, W.-P. Zhu, Tstnn: Two-stage transformer based neural network for speech enhancement in the time domain, in *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2021), pp. 7098–7102
24. E. Kim, H. Seo, Se-conformer: time-domain speech enhancement using conformer, in *Interspeech* (2021), pp. 2736–2740
25. Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, L. Xie, Dccrn: deep complex convolution recurrent network for phase-aware speech enhancement. arXiv preprint arXiv:2008.00264 (2020)
26. G. Yu, A. Li, C. Zheng, Y. Guo, Y. Wang, H. Wang, Dual-branch attention-in-attention transformer for single-channel speech enhancement, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 7847–7851
27. L. Wang, W. Wei, Y. Chen, Y. Hu, D2net: a denoising and dereverberation network based on two-branch encoder and dual-path transformer, in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (IEEE, 2022), pp. 1649–1654
28. D. Yin, C. Luo, Z. Xiong, W. Zeng, Phasen: a phase-and-harmonics-aware speech enhancement network, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34 (2020), pp. 9458–9465
29. D. Yin, Z. Zhao, C. Tang, Z. Xiong, C. Luo, Tridentse: guiding speech enhancement with 32 global tokens. arXiv preprint arXiv:2210.12995 (2022)
30. Y.-X. Lu, Y. Ai, Z.-H. Ling, Explicit estimation of magnitude and phase spectra in parallel for high-quality speech enhancement. arXiv preprint arXiv:2308.08926 (2023)
31. R. Cao, S. Abdulatif, B. Yang, Cmgan: conformer-based metric gan for speech enhancement. arXiv preprint arXiv:2203.15149 (2022)
32. S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, Y. Tsao, Metricgan+: an improved version of metricgan for speech enhancement. arXiv preprint arXiv:2104.03538 (2021)
33. N.L. Westhausen, B.T. Meyer, Dual-signal transformation lstm network for real-time noise suppression. arXiv preprint arXiv:2005.07551 (2020)
34. U. Isik, R. Giri, N. Phansalkar, J.-M. Valin, K. Helwani, A. Krishnaswamy, Poconet: better speech enhancement with frequency-positional embeddings, semi-supervised conversational data, and biased loss. arXiv preprint arXiv:2008.04470 (2020)

35. X. Hao, X. Su, R. Horaud, X. Li, 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)* (IEEE, 2021), pp. 6633–6637
36. A. Li, W. Liu, C. Zheng, C. Fan, X. Li, 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**, 1829–1843 (2021)
37. S. Zhao, B. Ma, K.N. Watcharasupat, W.-S. Gan, Frcrn: boosting feature representation using frequency recurrence for monaural speech enhancement, in *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 9281–9285
38. A. Li, C. Zheng, L. Zhang, X. Li, Glance and gaze: a collaborative learning framework for single-channel speech enhancement. Appl. Acoust. **187**, 108499 (2022)
39. A. Li, S. You, G. Yu, C. Zheng, X. Li, Taylor, can you hear me now? A taylor-unfolding framework for monaural speech enhancement. arXiv preprint arXiv:2205.00206 (2022)
40. L. Liu, H. Guan, J. Ma, W. Dai, G. Wang, S. Ding, A mask free neural network for monaural speech enhancement. arXiv preprint arXiv:2306.04286 (2023)

Publisher's Note