Spectral network based on lattice convolution and adversarial training for noise-robust speech super-resolution

Junkang Yang,[1] Hongqing Liu,[1,2,a] Lu Gan,[3] and Xiaorong Jing[2]

[1]*School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China*

[2]*Chongqing Key Lab of Mobile Communications Technology, Chongqing University of Posts and Telecommunications, Chongqing, China*

[3]*College of Engineering, Design and Physical Science, Brunel University, London, U.K.*

(Dated: 24 October 2024)

Speech super-resolution (SSR) aims to predict a high-resolution (HR) speech signal from its low-resolution (LR) counterpart. The previous models usually perform this task at a fixed sampling rate, reconstructing only high-frequency spectrogram components and merging them with low-frequency ones in noise-free cases. These methods achieve high accuracy but they are less effective in real-world settings, where ambient noise and flexible sampling rates are presented. To develop a robust model that fits practical applications, in this work, we introduce Super Denoise Net (SDNet), a neural network for noise-robust SR with flexible input sampling rates. To this end, SDNet's design includes gated and lattice convolution blocks for enhanced repair and temporal-spectral information capture. The frequency transform blocks are employed to model long frequency dependencies, and a multi-scale discriminator is proposed to facilitate the multi-adversarial loss training. The experiments show that SDNet outperforms current state-of-the-art noise-robust SSR models on multiple test sets, indicating its robustness and effectiveness in real-world scenarios.

[a]hongqingliu@outlook.com

I. INTRODUCTION

Speech super-resolution (SSR) aims to reconstruct missing high-frequency components from known low-resolution speech signals, thereby rendering speech clearer, more natural, and easier to comprehend. For speech communication, the application of SSR technology can effectively enhance call quality, mitigate distortion, and augment the intelligibility and comfort of speech. Furthermore, for numerous downstream tasks, this technique can also aid machines in better understanding human language, thus improving the accuracy and efficiency of tasks such as speech recognition (Haws and Cui, 2019) and speech synthesis (Yoneyama *et al.*, 2023).

Early SSR approaches primarily employed signal processing methods grounded in source-filter theory (Taylor and Reby, 2010), which models the speech signal as a product of the source signal passing through a vocal tract filter. These methods effectively extend the low-frequency signal by mapping low-frequency to high-frequency features using statistical techniques. However, with the advent and development of deep learning technology, neural network-based methods have emerged as the predominant approach in this field, demonstrating superior performance.

Despite substantial progress in deep learning-based SSR in recent years, current frequency domain-based approaches typically keep the low-frequency components and predict only the high-frequency components before combining the two. This method performs well in noise-free environments but fails to remove noise from the low-frequency part, introducing distortions in high-frequency prediction due to noise interference. As most existing net-

works are not designed to handle noise, retraining them with noisy data will be ineffective. Additionally, existing SSR models operate under fixed configurations, transforming specific low-sample-rate inputs to specific high-sample-rate outputs, and do not generalise well across different low-sampling rates. Performance degrades with varying speech databases or low-resolution signals generated by different downsampling schemes (Wang and Wang, 2021).

Recent studies have achieved high-quality speech super-resolution with flexible sampling rate inputs, often using Mel spectroscopy followed by neural vocoder synthesis. While these methods produce high-quality speech at 44.1 kHz or 48 kHz in noise-free environments, their large model sizes and numerous parameters complicate training and inference. Some efforts have focused on noise robustness in SSR, employing multi-stage training strategies or intermediate variable adjustments to jointly remove noise and increase the sampling rate from 8 kHz to 16 kHz. However, these methods suffer significant performance drops with certain noise types or low signal-to-noise ratios, and their complexity affects model's reproducibility.

For image super-resolution, convolutional neural networks with lattice blocks have demonstrated remarkable superiority (Luo *et al.*, 2022, 2020a), and this design has been successfully applied to more complex image restoration tasks. Despite their success in computer vision, these networks have not yet been explored for audio restoration tasks, including SSR, noise suppression, and packet loss concealment.

To improve SSR performance in noisy environments, we propose Super Denoise Net (SD-Net), a neural network designed to remove noise while extending bandwidth. Drawing

inspiration from image restoration, we incorporate gated convolution to boost the network's generative capability and introduce lattice convolution blocks in the bottleneck layer to capture more information in the time-frequency domain. Training the model with large, noise-containing datasets, our experiments show that SDNet significantly outperforms existing SSR models in both objective and subjective evaluations. An ablation study further highlights the impact of our design on model performance.

Our main contributions are summarised as follows:

- We introduced lattice blocks and gated convolution structures, which have proven effective in image restoration, to the SSR task, enhancing the network's recovery capabilities.
- Through data augmentation, we achieved greater noise robustness compared to existing noise-robust speech super-resolution models, without requiring prior knowledge of the input signal's sampling rate.
- Within an adversarial training framework, we developed a multi-scale discriminator strategy to optimise multiple loss functions
- Our model outperforms the baseline in both noise-free and noisy environments, employing a simpler training strategy and resulting in negligible artifacts in the transition frequency band.

The rest of the paper is organized as follows. In Section II, we introduce the settings of the task addressed in this article and its related works. In Section III, we describe the details of the proposed network and the data processing method. We document the details

of the experimental settings as well as the different baselines in Section IV. In Section V, we report and analyse the experimental results, followed by the conclusion and future works in Section VI.

II. PROBLEM FORMULATION AND RELATED WORKS

A. Modeling of SSR Task

Speech super-resolution is also known as the bandwidth extension of speech signals in many previous works. In the time domain, low-sampling-rate speech contains fewer sample points for the same duration, and the super-resolution model predicts extra sample points based on the information from the low-sampling-rate speech waveform, so that they are converted into high-sampling-rate speech with better sound quality. From the aspect of frequency domain, due to the increase in the number of sampling points in the same duration, the missing high-frequency portion of the low-sampling-rate speech signal is supplemented.

In a formal setting, noted in previous work's description (Kuleshov *et al.*, 2017), we represent a low-resolution speech waveform as $x(t), t = 0, 1, ..., T, T \in \mathbb{R}$, where T is the duration (in seconds) of this signal and $x(t)$ is the amplitude at time t. When it is sampled at a sampling rate of R_1 Hz, $t = \frac{1}{R_1}, \frac{2}{R_1}, ..., T$, and the goal of SSR is to generate a high-resolution version $\hat{y}(t)$ of $x(t)$ that has a sampling rate $R_2 > R_1$ and the same duration as $x(t)$, where $t = \frac{1}{R_2}, \frac{2}{R_2}, ..., T$.

In noisy environments, the speech signal is corrupted by noises, which can be expressed by

$$x_n(t) = x(t) + n(t), \tag{1}$$

where $n(t)$ is the noise with a sampling rate of R_1 Hz. The noise-robustness of SSR means the capability to restore the high-resolution clean version $\hat{y}_c(t)$ of $x_n(t)$, i.e. removing the noise of low-frequency part and predicting the clean high-frequency part at the same time.

B. Noise-Free SSR Methods

Most early SSR approaches are bandwidth extension (BWE) methods based on traditional signal processing theory and the source-filter speech generation model (Taylor and Reby, 2010). Within this framework, various techniques have been developed to estimate wide-band spectral envelopes (Cheng *et al.*, 1994; Park and Kim, 2000), including methods based on Gaussian Mixture Models (GMM) (Nour-Eldin and Kabal, 2009), Hidden Markov Models (HMM) (Bauer and Fingscheidt, 2008), and codebook mapping (Pulakka *et al.*, 2013).

With the current developments of deep learning, new methods and models to further improve the performance of SSR tasks have been proposed (Birnbaum *et al.*, 2019; Li and Lee, 2015; Ling *et al.*, 2018; Wang and Wang, 2021). TFNet (Lim *et al.*, 2018) enhances the SSR quality by jointly optimizing both the time and frequency domain. AFiLM (Rakotonirina, 2021) introduces self-attention based on TFiLM (Kuleshov *et al.*, 2017) and achieves a better performance with a faster inference speed. Utilizing a U-Net, (Li *et al.*, 2021) and (Nguyen *et al.*, 2022) improve the SSR accuracy under a constraint of low complexity, with pre-training and self-supervised learning methods, respectively.

These previous studies have been centered on transforming the speech signal from narrow-band to wide-band, and super-resolution to a higher resolutions was still not achieved.

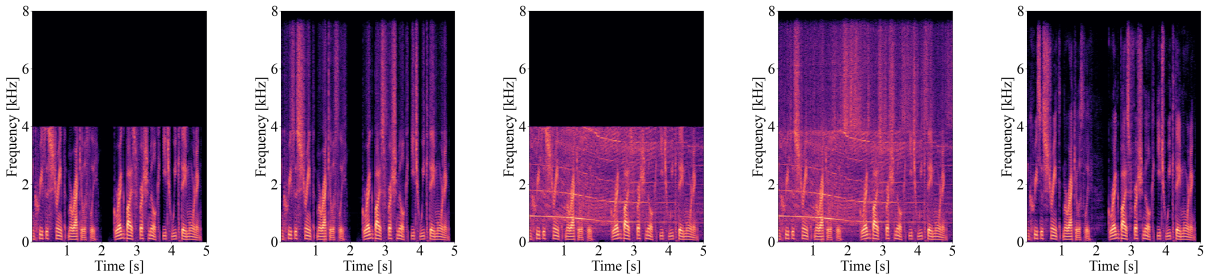

FIG. 1. (color online) Spectrograms of reconstructed and original speech. (a) Narrow-band speech without noise. (b) Wide-band version of (a) generated by a noise-free SSR model. (c) Narrow-band speech containing noise. (d) Wide-band version of (c) generated by a noise-free SSR model. (e) Wide-band version of (c) generated by a noise-robust SSR model.

With the background of a general promotion in the quality of network communications, recent works mainly focus on generating high-fidelity and full-band speech (Zhang *et al.*, 2021). As has been verified in computer vision, generative model has high the potential in generative tasks like image super-resolution and reconstruction, so generative adversarial network (GAN) and diffusion based methods are widely adopted in current SSR works (Han and Lee, 2022; Moliner and Välimäki, 2023; Shuai *et al.*, 2023; Yoneyama *et al.*, 2023; Yu *et al.*, 2023). In (Mandel *et al.*, 2023), the authors propose a GAN operating in the frequency domain to eliminate the artifacts at the transition region between existing and generated frequency bands. BAE-Net (Yu *et al.*, 2024) addresses the fluctuations of effective bandwidth in real-world audio for SSR. With a latent diffusion model and a neural vocoder, AudioSR (Liu *et al.*, 2024) handles the super-resolution of speech, music recording, and sound effects. A similar two-stage vocoder-based structure was also used by Fre-Painter (Kim *et al.*, 2024).

However, many of these studies are limited by fixed sampling rates and the concatenation of different bands, leading to degraded performance in noisy environments (see FIG. 1(d) and artefacts in transition parts (see FIG. 1(a). These issues also hinder retraining the original models with noisy data.

C. Noise-Robust SSR Methods

In practical scenarios, speech signals are often corrupted different noises and present various bandwidth ranges, which makes it hard to directly improve their quality by most noise-free SSR models, and this issue is visualized in Figure 1. In order to make SSR techniques more practical, it is crucial to investigate the robustness and bandwidth-adaption in complex environments.

A typical approach to solve the noise problem in SSR task is to first perform speech enhancement on noisy narrow-band signals and then followed by a bandwidth extension under noise-free conditions. For example, (Moreno *et al.*, 1996) applied an iterative vector Taylor series (VTS) approximation algorithm on feature enhancement, and then reconstruct the wide-band signal with a Gaussian mixture model or a maximum a posterior estimation (Seltzer *et al.*, 2005; Seo *et al.*, 2014). The same approach also applies to two-stage neural network (Chen *et al.*, 2022; Liu *et al.*, 2018; Taher *et al.*, 2023). These methods, although simple and straightforward, faces difficulties in phase estimation. In addition, some multi-task models (Hernandez-Olivan *et al.*, 2024; Moliner *et al.*, 2023) consider noise, clipping, and bandwidth loss simultaneously, but such approaches deal with different single tasks

separately with a versatile framework and are not effective for the case where multiple distortions co-exist.

To further develop the noise-robust SSR model, (Hou *et al.*, 2020) proposes a multitasking framework that reconstructs clean wide-band signals directly from noisy narrow-band signals by introducing intermediate variables into the loss function. VoiceFixer (Liu *et al.*, 2022) fixes multiple distortions simultaneously in the Mel-spectrum domain and then reconstructs the waveform with a neural vocoder. In 2023, (Lin *et al.*, 2023) proposed EP-WUN based on the WaveUNet backbone (Stoller *et al.*, 2018). To treat noise suppression and super-resolution jointly, the method uses three stages of training and introduces intermediate variables into the improved triplet loss. The authors claim that the model achieves the state-of-the-art performance on noise-robust SSR task currently and introduce a large positive impact on the accuracy of the speech recognition task.

In summary, there are relatively few studies on noise-robust SSR compared to noise-free SSR. Most models extend 8 kHz recordings to 16 kHz for clean speech, leaving room for improvement in robust bandwidth adaptation. Additionally, there is a need for the development of simple and effective training algorithms.

III. PROPOSED NETWORK

Figure 2 illustrates the SSR model proposed in this paper. It employs a GAN architecture comprising pre- and post-processing modules, a generator operating in the spectral domain, and a multi-scale discriminator. Initially, we perform a short-time Fourier transform (STFT) on the narrow-band speech and obtain the wide-band speech by zero-padding

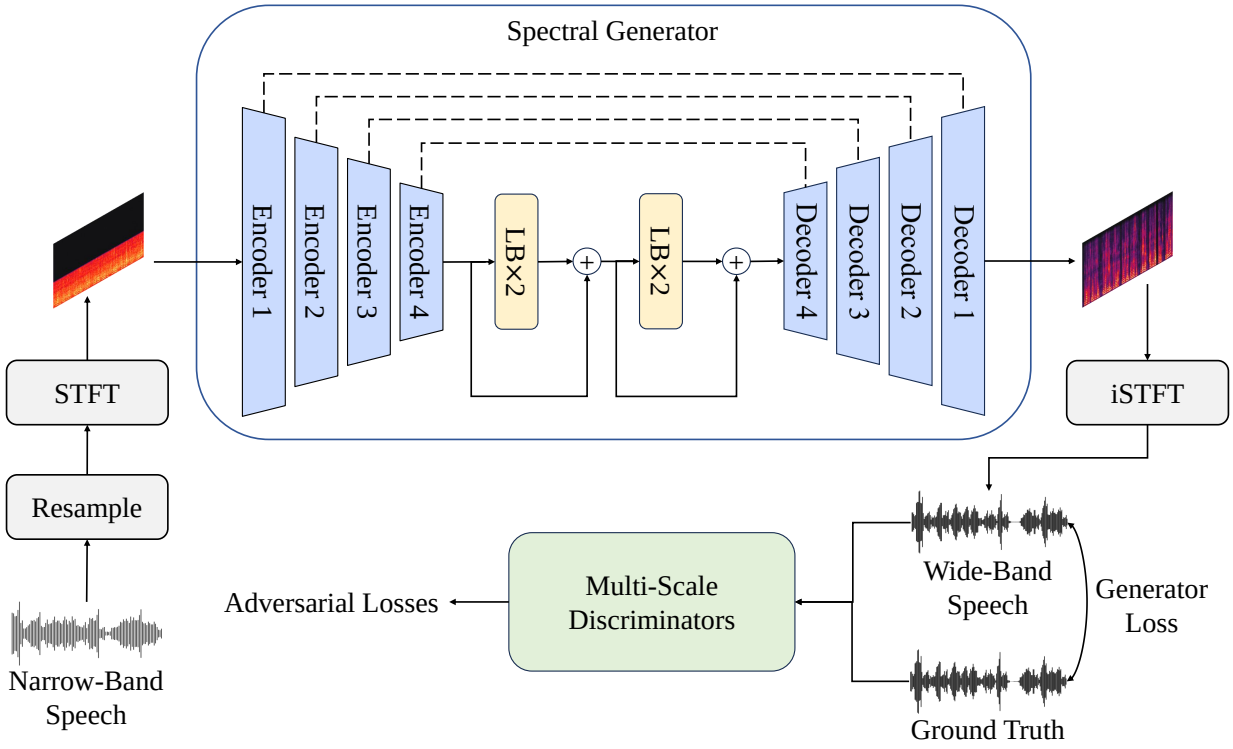

FIG. 2. (color online) The structure of proposed generative adversarial network.

the high-frequency part through resampling. Unlike conventional noise-free methods, our resampling step maintains the same scale for input and output, enabling the network to make comprehensive end-to-end predictions across the entire bandwidth, thus overcoming the limitations of previous splicing methods that fail to eliminate low-frequency noise and produce artefacts. The generator features a traditional U-shaped structure with encoder and decoder modules and a bottleneck layer. Notably, we incorporate lattice convolution blocks (LBs) in the bottleneck layer to capture both local and global dependencies effectively, reducing computational complexity while preserving modelling capability through sparse connectivity. Detailed descriptions of each module are given below:

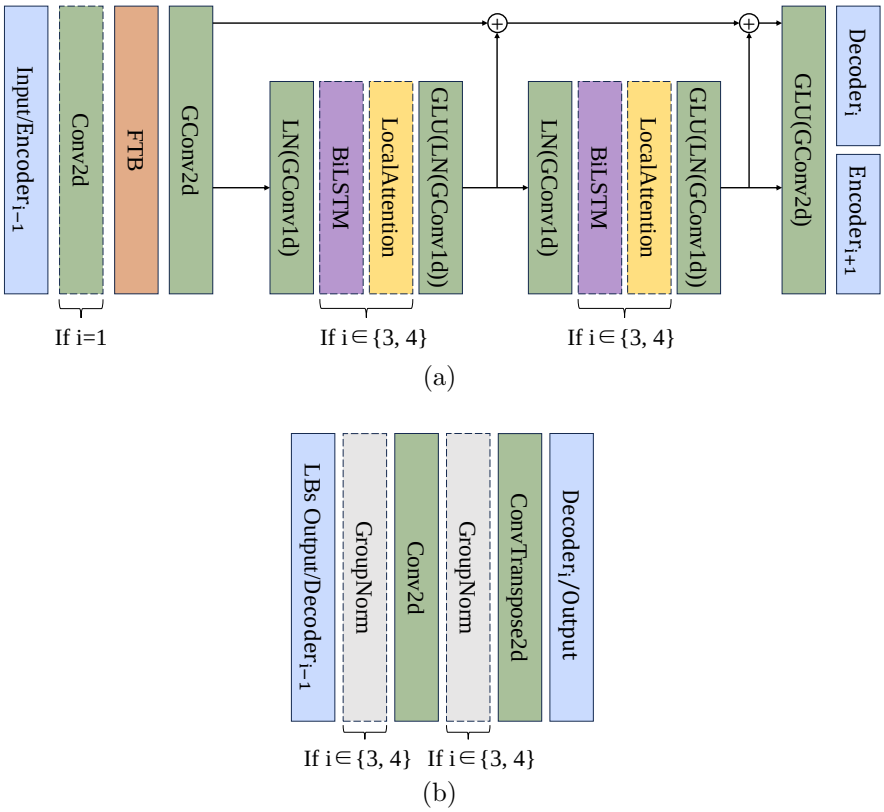

FIG. 3. (color online) The structure of encoder layer (a) and decoder layer (b).

A. Spectral Generator

1. Encoder and Decoder Layer

As depicted in FIG. 2, the encoder-decoder framework consists of four layers each, facilitating the transformation of input data into a latent representation and its subsequent reconstruction. With the encoder, depicted in FIG. 3(a), the initial layer undertakes the reshaping of the input via a 2D convolution operation. Following this, a pivotal frequency transform block (FTB) (Yin *et al.*, 2020) intervenes to capture non-local correlations within the spectrogram, traversing along the frequency axis. The operations within each FTB can be succinctly represented by a distinct formula, indicated in FIG. 4(a):

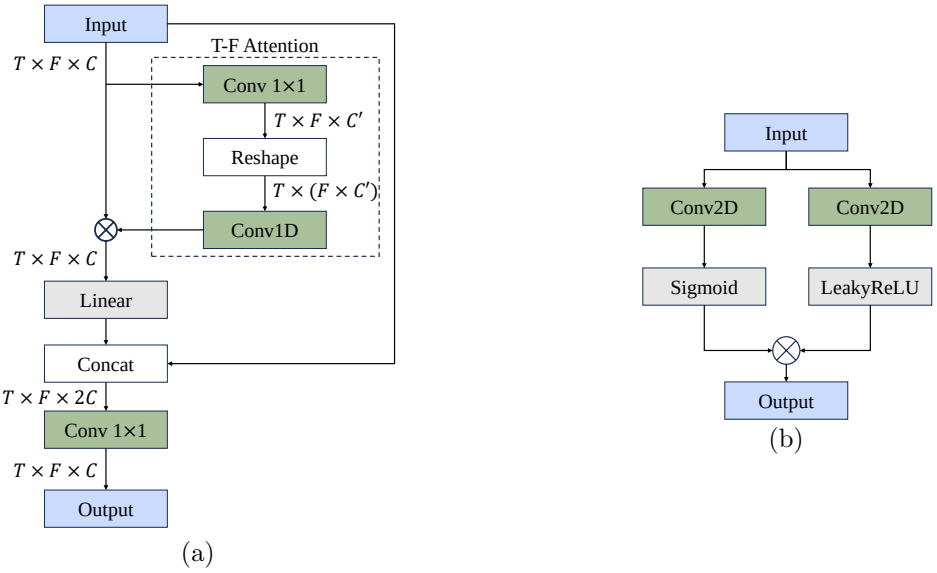

FIG. 4. (color online) The structure of FTB module (a) and gated convolution (b).

$$\boldsymbol{X_O} = Conv(Concat(Linear(\boldsymbol{X_I} \otimes Attn(\boldsymbol{X_I})))), \tag{2}$$

where $\boldsymbol{X_I}$, $\boldsymbol{X_O}$, and $\otimes$ represent input, output spectrogram tensor, and point-wise multiplication, with the $Attn(\cdot)$ operation highlighted in the dotted box. In the context of time-frequency domain, non-local correlations manifest along the frequency axis. A prominent example of such correlations pertains to harmonics, which have been demonstrated to aid in reconstructing distorted spectrograms. However, a direct concatenation of 2D convolution layers with small kernels fails to adequately capture these global correlations. In this work, we set the incorporation of FTBs at the beginning of the residual branches to address this limitation, ensuring that the resulting features encompass a comprehensive frequency receptive field. The gated convolution (GConv) was first proposed in free-form image inpainting (Yu *et al.*, 2019), which has similar points with SSR task in uncertain sampling rates. In FIG. 4(b), with a soft-masking and a featuring branch, gated convolution layer learns a dynamic feature selection mechanism for each channel and each spatial

location, promoting the adaption of our model for various bandwidth distortion in complex environments. It is formulated by

$$\boldsymbol{M} = \sum \sum \boldsymbol{W_{MC}} \cdot \boldsymbol{X_{in}}, \tag{3}$$

$$\boldsymbol{F} = \sum \sum \boldsymbol{W_{FC}} \cdot \boldsymbol{X_{in}}, \tag{4}$$

$$\boldsymbol{X_{out}} = \phi(\boldsymbol{F}) \otimes \sigma(\boldsymbol{M}), \tag{5}$$

where σ is sigmoid function and ϕ can be any activation functions. In our study, ϕ is LeakyReLU (slope=0.2, inplace=True), $\boldsymbol{W_{MC}}$ and $\boldsymbol{W_{FC}}$ are convolutional filters of soft-masking and feature branches, $\boldsymbol{M}$ and $\boldsymbol{F}$ denote mask and feature tensor.

Within the inner encoder architecture, a dual residual branch is employed, with the insertion of two 1D gated convolutions at the ingress and egress points. Situated centrally, crucial components including bidirectional long short-term memory (BiLSTM) units and attention module serve to model long-range dependencies, enriching the model's capacity to discern temporal correlations across the spectral latent space. Sequentially, each encoder layer is succeeded by a corresponding decoder layer (see FIG. 3(b)), which aims to reconstruct latent vectors commensurate with the spectrogram's dimensions after they passed through the encoder layers. Specifically, a concatenated residual connection is set between each encoder and decoder layer, facilitating the seamless flow of information across the encoding-decoding. Conversely, within the encoder layer, a summative residual connection between two residual branches is instantiated, consolidating information flow and mitigating the vanishing gradient phenomenon across the network. These architectural designs collectively

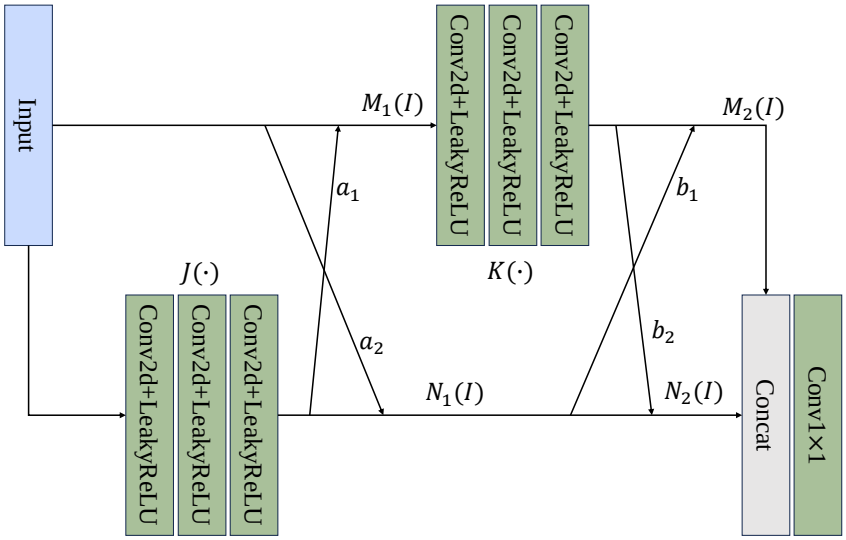

FIG. 5. (color online) The structure of lattice block.

contribute to the model's efficacy in capturing intricate spectral dependencies intrinsic to the noisy speech data.

2. Lattice Convolution Blocks

The bottleneck layers of our model include four lattice convolution blocks (LBs), a novel concept initially introduced in the domain of image restoration tasks (Luo *et al.*, 2022, 2020a). When integrated with gated convolution, this architectural arrangement presents a fusion of structured interpolation alongside adaptive and expansive context modeling capabilities. As depicted in FIG. 5, each LB module consists of paired lattice structures. Input data traverses through two distinct branches, each comprising multiple convolutional layers, with a subsequent LeakyReLU activation layer following each convolutional operation. Notably, these two branches engage in mutual interaction facilitated by learnable combination coefficients, fostering collaborative feature extraction and representation. Specifically, given

an input feature I, the first combination is

$$M_1(I) = I + a_1 J(I), \tag{6}$$

$$N_1(I) = a_2 I + J(I), \tag{7}$$

where $J(\cdot)$ denotes to the implicit non-linear function of several layers shown in FIG. **??**. Similarly, the second combination is

$$M_2(I) = b_1 N_1 + K(M_1(I)), \tag{8}$$

$$N_2(I) = N_1 + b_2 K(M_1(I)). \tag{9}$$

Then, the outputs of two branches are merged in channel dimension and then compressed by a 1×1 convolution layer. The final output is

$$O = Conv(Concat(M_2(I), N_2(I))). \tag{10}$$

The combination coefficients are mainly determined in the following way. The mean and standard deviation in channel dimension are first obtained by global mean pooling in the upper branch and global standard deviation pooling in the lower branch. Then, those statistics in two branches are passed through two fully connected layers, each followed by ReLU and Sigmoid activation functions, respectively. Finally, the outputs of the two branches are averaged to obtain the combined coefficients.

B. Multi-Scale Discriminator

To implement multi-loss training in an adversarial framework for enhancing SSR speech quality, we leverage multi-scale discriminators, illustrated in FIG. 6. These discriminators are integral components of the system, analyzing inputs comprising SR speech and

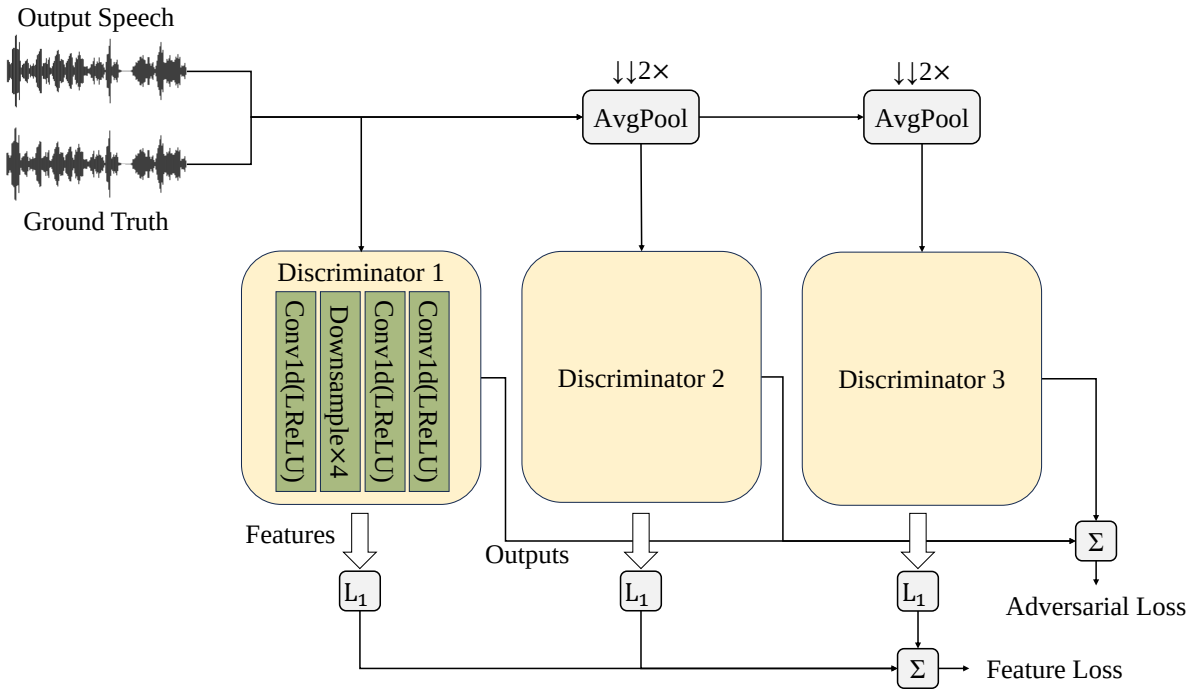

FIG. 6. (color online) The multi-scale discriminators architecture.

high-resolution reference signals, both synthesized by the generator. Comprising a trio of discriminators denoted by D_1, D_2, D_3, each adheres to the structural design in MelGAN. In particular, each discriminator consists of 7 convolutional layers, with 4 layers equipped with downsampling capabilities. As data traverse through these layers, they yield real and fake features across distinct scales, pivotal for computing the feature loss. Additionally, the discriminator's outputs contribute to the computation of adversarial losses for both the generator and discriminator. Moreover, it is worth highlighting that the inputs provided to D_1, D_2, and D_3 represent original waveforms, 2-times down-sampled waveforms, and 4-times down-sampled waveforms, respectively, augmenting the discriminators' capability to discern features at varied resolutions. For in-depth insights into the discriminators' architectural specifics, readers are encouraged to refer to (Kumar *et al.*, 2019).

C. Loss Function

The model is trained with an adversarial approach. We use a multi-scale STFT loss with FFT bins $\in \{512, 1024, 2048\}$ and hop length $\in \{50, 120, 240\}$ to form one part of the loss function. The window lengths are $\{240, 600, 1200\}$. On the other hand, the multi-scale adversarial and feature losses in the time domain are also added in. The total loss is

$$\mathcal{L} = \mathcal{L}_{MSTFT} + \mathcal{L}_G^{adv} + \lambda_f \mathcal{L}_f, \tag{11}$$

where $\lambda_f = 100$, $\mathcal{L}_{MSTFT}$, $\mathcal{L}_G^{adv}$ and $\mathcal{L}_f$ are multi-scale STFT loss, adversarial loss of generator, and feature loss, respectively. Let $s(x, \theta_m)$ denote $|STFT(x)|$ with the m-th hyperparameters θ_m, the multi-scale STFT loss is defined as

$$\begin{aligned} &\mathcal{L}_{MSTFT} = \mathbb{E}_{(x,y)\sim p_{data}} \\ &\left[\sum_{m=1}^{3}\left(\frac{||s(y,\theta_m) - s(x,\theta_m)||_F}{||s(y,\theta_m)||_F} + \frac{1}{N}||log\frac{s(y,\theta_m)}{s(x,\theta_m)}||\right)\right], \end{aligned} \tag{12}$$

where $||\cdot||_F$ and $||\cdot||_1$ are Frobenius and ℓ_1-norms, N is the number of elements in the magnitude.

As shown in FIG. 6, the latter two loss functions can be depicted as

$$\mathcal{L}_G^{adv} = \mathbb{E}_{x\sim p_{data}}\left[\frac{1}{K}\sum_k max(0, 1 - D_k(G(x)))\right], \tag{13}$$

$$\mathcal{L}_f = \mathbb{E}_{(x,y)\sim p_{data}}\left[\frac{1}{KL}\sum_{k,l}||D_k^l(y) - D_k^l(G(x))||_1\right], \tag{14}$$

where $k = 1, ..., K$ is the number of discriminators, $l = 1, ..., L$ is the number of layer in one discriminator.

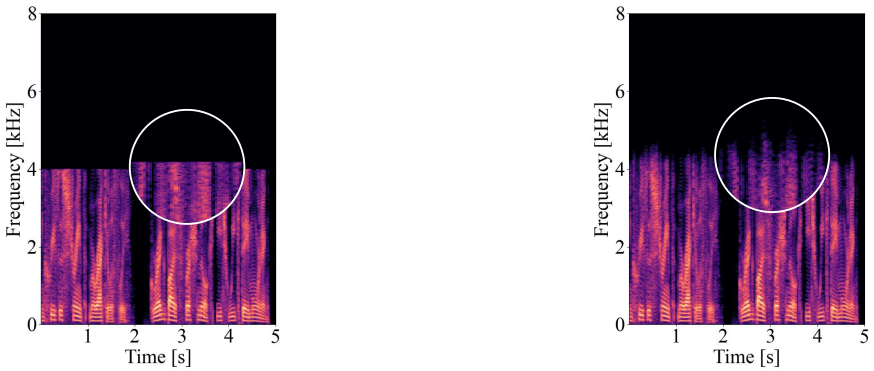

FIG. 7. (color online) Spectrograms of the downsampled speech using resample function (a) and our proposed method (b).

IV. EXPERIMENTS

A. Data Augmentation

In this study, we use the dataset from the Deep Noise Suppression (DNS) Challenge presented at ICASSP 2023 (Dubey *et al.*, 2024) and the corpus compiled by Valentini-Botinhao (Valentini-Botinhao *et al.*, 2016). This combination provides comprehensive training data with diverse noise profiles, representative of real-world scenarios. We constructed the training data by synthesising clean and noisy speech pairs through random mixing of speech and noise components, resulting in a 500-hour audio dataset. Each sample is standardised to 5 seconds, with controlled signal-to-noise ratios (SNR) ranging from -5 dB to 20 dB to mimic real-world conditions. Additionally, 16 kHz sampling rate is applied for all samples, ensuring compatibility with contemporary audio processing frameworks

In most existing SSR training dataset generation, a fixed filter with a fixed sampling rate or a direct resampling function is used (Xu *et al.*, 2023), leading to the artefacts in

ALGORITHM 1: Downsampling algorithm for flexible sampling rates cases.

Data: $\boldsymbol{y} \in \mathbb{Y}$
Result: The high-quality speech $\boldsymbol{y}$ and its downsampled version x
$\boldsymbol{x} = \boldsymbol{s}$;
$type = randomType\ (Chebyshev, Elliptic, Butterworth, Boxcar)$;
$f_{cut} \sim U(C_{low}, C_{high})$;
$order \sim U(O_{low}, O_{high})$;
$\boldsymbol{x} = \boldsymbol{x} * Filter(type, f_{cut}, order)$;
if resample **then**
  $\boldsymbol{x} = Resample(Resample(\boldsymbol{x}, 16000, f_{cut} \times 2), f_{cut} \times 2, 16000)$;
end if

the generated spectrogram. Inspired by (Liu *et al.*, 2022), we employ a filtering mechanism with stochastic parameters. As outlined in Algorithm 1, the filter types include *Chebyshev*, *Elliptic*, *Butterworth*, and *Boxcar*, each offering distinct characteristics. The filter order is determined by a randomly generated integer ranging from 2 to 10, ensuring variability and robustness. The cutoff frequency, essential for defining the filter's behaviour, ranges from 2 kHz to 8 kHz, covering a bandwidth relevant to our study. This data augmentation strategy preserves the transition region between high and low-frequency bands, as shown in Figure 7. Besides, by leveraging varied filters and downsampling factors, our approach prevents the model from being overly tailored to any single type of filtering or downsampling process. This diversity in training conditions equips the model to perform well across a wider range of real-world scenarios, enhancing its generalisation capability.

B. Implementation Settings

In contrast to the complex training strategies in prior research, which often involve multi-stage training, variable learning rates, and warm-up procedures, our proposed methodology adopts a streamlined, single-stage approach. Specifically, we use the Adam optimizer with parameters $\beta_1 = 0.8$ and $\beta_2 = 0.999$, maintaining a consistent learning rate of 1×10^{-4} for

both the generator and discriminator components. Training spans 200 epochs on NVIDIA RTX3090 GPUs. To evaluate the model's generalisation and performance, we use a validation dataset and select the checkpoint from the epoch with the best performance for further testing. This protocol aims to create a robust and straightforward pipeline yielding promising results across varied evaluation metrics. For more details on our parameter setup, please refer to our demo page[1].

C. Baselines

For the noise-free SSR task, we selected the following baselines:

- WSRGlow (Zhang *et al.*, 2021): It combines glow model and WaveNet (Stoller *et al.*, 2018) for audio super-resolution, introducing LR and STFT encoders to generate full-band audio.
- NU-Wave 2 (Han and Lee, 2022): NU-Wave 2 is a diffusion model that generates high-quality 48 kHz audio from various input sampling rates using fewer parameters.
- VoiceFixer (Liu *et al.*, 2022): VoiceFixer is a two-stage neural vocoder framework designed for general speech restoration, capable of handling multiple distortions like denoising, dereverberation, super-resolution, and declipping in a unified model.
- AERO (Mandel *et al.*, 2023): It is a spectral-domain GAN based model using a U-Net generator to predict high-frequency content, surpassing state-of-the-art methods.
- AudioSR (Liu *et al.*, 2024): AudioSR is a diffusion-based model for versatile audio super-resolution which can handle various audio types (speech, music and sound ef-

fects) to in Mel domain and using a HiFi-GAN (Kong *et al.*, 2020) neural vocoder to generate audio waveforms.

However, some of these do not support flexible input sampling rates, so comparisons were made only with Nu-Wave 2, VoiceFixer, and AudioSR when the sampling rate was flexible.

For the noise-robust SSR task, we compared our model with the previous state-of-the-art methods:

- UEE (Liu *et al.*, 2018): This is a unified framework for speech enhancement and bandwidth extension using jointly trained BLSTM-RNNs, with multi-task transfer learning for model compression.
- MTL-MBE (Hou *et al.*, 2020): It is a noise-robust bandwidth extension framework using multi-task learning and time-domain masking for joint speech enhancement and bandwidth extension.
- EP-WUN (Lin *et al.*, 2023): EP-WUN is a noise-robust bandwidth extension model that enhances Wave-U-Net (Stoller *et al.*, 2018) with a speech quality classifier and a modified triplet loss to improve speech representation for 8 kHz speech.
- I-DTLN + AFiLM (Chen *et al.*, 2022): The proposed model integrates Unet+AFiLM and I-DTLN to create a system for audio super-resolution and noise cancellation in low sampling rate and noisy environments.

As the authors did not provide source code, we re-implemented the method proposed in (Chen *et al.*, 2022) to produce the results. For uncertain input sampling rates, VoiceFixer was used as the baseline, being the only model currently supporting this case.

To evaluate the denoising performance of our method, we conducted a 16 kHz to 16 kHz denoise task on the same dataset using various popular denoise neural models, including:

- TSTNN (Wang *et al.*, 2021): A two-stage transformer-based neural network for time-domain speech enhancement.
- DPRNN (Luo *et al.*, 2020b): A dual-path recurrent neural network, splitting input sequences into chunks for local and global processing.
- TFT-Net (Tang *et al.*, 2020): A cross-domain speech enhancement model that uses a dual-path attention block to enhance spectrogram-to-waveform conversion.
- DCCRN (Hu *et al.*, 2020): A deep complex convolution recurrent network for phase-aware speech enhancement using complex CNNs and LSTMs.
- FullSubNet (Hao *et al.*, 2021): A real-time speech enhancement model that fuses full-band and sub-band information to capture both global spectral context and local signal details.
- DPT-FSNet (Dang *et al.*, 2022): A dual-path transformer-based network that fuses full-band and sub-band information for improved speech enhancement in the frequency domain.

The baseline system for all experiments in this article is from the above model. For all the baselines we follow the original settings and they will be re-trained in our experiments if necessary. The audio clips are stored in '.wav' format with 16 bit depth unless otherwise specified.

366 ### D. Objective Evaluation Metrics

367 The following objective evaluation metrics are employed:

368 **Perceptual evaluation of speech quality (PESQ)** (Rix *et al.*, 2001). PESQ is a
369 metric for assessing speech quality, with a range from -0.5 to 4.5. The closer the value is to
370 this upper limit, the higher quality the speech has. It also has two versions, including both
371 narrow-band (PESQ-NB, 0-8 kHz) and wide-band (PESQ-WB, 8-16 kHz).

372 **Short-Time Objective Intelligibility (STOI)** (Taal *et al.*, 2011). STOI evaluates
373 the objective intelligibility of a degraded speech signal by computing the correlation of the
374 temporal envelopes of the degraded speech signal and its clean reference (Zhao *et al.*, 2024).
375 It ranges from 0 to 1 and the higher value represents the better quality.

376 **CSIG, CBAK and COVL** (Hu and Loizou, 2007). The CSIG, CBAK, and COVL
377 are the **Mean Opinion Score (MOS)** prediction of signal distortion, intrusiveness of
378 background noise, and overall effect, and they all range from 0 to 5. CSIG predicts the
379 rating of speech distortion. Higher CSIG values indicate better performance in reducing
380 distortion. CBAK evaluates the intrusiveness of background noise distortion. Higher CBAK
381 values indicate better noise suppression. COVL combines CSIG and CBAK to provide an
382 overall score of processed speech quality.

383 **Log Spectral Distance (LSD)**, which is defined by

384
$$S = 10\log_{10}|s(t,k)|^2,$$

385
$$\hat{S} = 10\log_{10}|s(t,k)|^2,$$
$$LSD(\hat{S},S) = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{1}{K}\sum_{k=1}^{K}(S-\hat{S})^2}, \tag{15}$$

where $s(t,k)$ and $s(\hat{t},k)$ represent the spectrogram of the ground truth and the reconstructed speech respectively, T, K denote the number of time frames and bins in the spectrogram. LSD is a distance measure between two spectra so the lower LSD means the produced speech has more similarity to the ground truth.

E. Subjective Evaluation Metrics

The subjective evaluation metric is overall MOS (835, 2003), which is a widely used metric for evaluating speech quality. It provides a subjective assessment of how well a listener perceives the quality of a speech signal. We randomly selected 50 samples from the each test set and asked 15 people to provide overall MOS score of each sample in a range of 5 levels (see table I). These listeners are native speakers and represent the target audience. The final MOS score is the average of these evaluations.

TABLE I. Levels of MOS score.

Score	Description
5	Excellent (Near-perfect quality.)
4	Good (Clear and pleasant.)
3	Fair (Acceptable but not ideal.)
2	Poor (Noticeably degraded.)
1	Bad (Unintelligible or severely distorted.)

V. RESULTS AND ANALYSIS

A. Noise-Free Cases

For noise-free cases, we first conducted comparison on DNS no-reverb test set with fixed sampling rate, i.e. 8 kHz to 16 kHz super-resolution, and the results are in Table II. We compare five state-of-the-art SSR methods using their original implementations alongside the proposed method as follows.

TABLE II. Test results of noise-free 8k to 16k SSR task on DNS no-reverb test set.

Method	PESQ-NB	PESQ-WB	STOI(%)	CSIG	CBAK	COVL	LSD	MOS
WSRGlow (Zhang *et al.*, 2021)	4.365	2.811	**99.4**	3.946	4.068	3.433	0.929	4.21
NU-Wave 2 (Han and Lee, 2022)	4.353	2.646	**99.4**	3.663	2.869	3.209	1.328	4.08
VoiceFixer (Liu *et al.*, 2022)	2.999	1.983	85.9	2.937	2.095	2.416	1.140	4.18
AERO (Mandel *et al.*, 2023)	4.369	3.295	98.5	**4.287**	4.273	3.844	0.802	4.27
AudioSR (Liu *et al.*, 2024)	4.368	2.299	98.8	3.464	2.952	2.937	1.141	4.29
Ours	**4.377**	**3.611**	98.6	4.103	**4.553**	**3.935**	**0.783**	**4.55**

In noise-free cases, our method presents significant advantages compared to other deep SSR baselines, with better performance in most metrics. In particular, in wide-band PESQ, our method outperforms the best baseline model by 0.316, which indicates that our method substantially improves speech quality over the entire bandwidth range and is not limited to the original or generated part. In CBAK, our method reaches 4.553, outperforming the baseline model by 0.28, which demonstrates that our special network design and data simulation methods for background noise are very effective. The performance of our method in CSIG is slightly lower than that of AERO. This may be due to the slight impairment of the speech component when removing noise, which is a common issue for all denoising neural models. However, our method is still the best performing one among the methods in the table due to overall sights.

Table III illustrates the test results when the sampling rates of input data are flexible, and the test set also do not contain noises. For this cases, all speech clips in test set were downsampled using the method proposed in Sec.IV, with the random sampling rates from 4 kHz to 16 kHz. Due to the random downsampling, the sampling rates of the narrow-band speech data is overall higher than that in Table 1, which also causes the objective metrics to be increased. Similarly, the number of baseline models under this experiment setup drops because most models do not support inference for data with flexible sampling rates. In a comparison with all baseline models, our method performs best across all objective metrics. Our model improves over the baseline by 0.974 on the broadband PESQ, and we achieve a performance of 4 or more in the CSIG, CBAK, and COVL metrics, which measure the effectiveness of the proposed method. The improvement in the objective metrics indicates

that the reconstructed speech using our method without prior sampling rates knowledge is already of high quality, and even for some narrow-band speech containing fewer distortion, and the reconstruction is very close to the ground truth.

TABLE III. Test results of noise-free SSR task with uncertain input sampling rates on DNS no-reverb test set.

Method	PESQ-NB	PESQ-WB	STOI(%)	CSIG	CBAK	COVL	LSD	MOS
NU-Wave 2 (Han and Lee, 2022)	4.397	3.407	98.4	4.102	3.274	3.819	1.193	4.23
VoiceFixer (Liu *et al.*, 2022)	2.974	2.179	85.9	3.191	2.191	2.641	1.086	4.21
AudioSR (Liu *et al.*, 2024)	4.262	2.911	98.0	3.920	3.271	3.504	1.012	4.36
Ours	**4.436**	**3.885**	**99.1**	**4.151**	**4.633**	**4.075**	**0.695**	**4.59**

B. Noise-Robust Cases

Table IV comprehensively compares our proposed model with existing deep noise-robust SSR methods, including the SOTA models. From the table, our model consistently outperforms other methods in most metrics, which validates its effectiveness in handling the joint task of SSR and noise suppression. In particular, for PESQ-WB and COVL, we observe excellent performance, ahead of the current SOTA method by 0.13 and 0.06, respectively.

TABLE IV. Test results of noise-robust SSR tasks. The 8 kHz source speeches are from Valentini-Botinhao noisy test set and the 4-16 kHz source speeches are from DNS no-reverb noisy test set.

Method	Source	PESQ-NB	PESQ-WB	STOI(%)	CSIG	CBAK	COVL	LSD
UEE (Liu *et al.*, 2018)	8 kHz	——	2.23	93	2.27	2.39	2.17	2.72
MTL-MBE (Hou *et al.*, 2020)		——	2.55	94	2.64	3.21	2.46	2.29
EP-WUN (Lin *et al.*, 2023)		——	2.25	92	**3.50**	2.94	2.86	1.23
AFiLM + I-DTLN (Chen *et al.*, 2022)		——	2.54	90	2.63	2.87	2.18	1.54
Ours		——	**2.67**	**95**	3.29	**3.32**	**2.92**	**1.16**
VoiceFixer (Liu *et al.*, 2022)	4-16 kHz	2.54	1.82	84.2	2.74	1.98	2.22	1.28
Ours		**3.55**	**3.01**	**97.3**	**3.66**	**3.73**	**3.36**	**1.11**

These results are in line with our initial expectations, verifying that our improvements to the network architecture and the use of novel simulations for the data not only improve the quality of the reconstructed high-frequency part, but also suppress the noise to a better extent. However, it is worth noting that our method exhibits a slight degradation in the

CSIG metric. This is because the suppression of noise, although beneficial to the overall quality, may unintentionally affect the speech parts, as we mentioned before. In conclusion, in addition to CSIG, our SDNet shows significant promise in noise-robust SSR. These findings highlight that our model is a balanced approach that optimises both noise reduction and SSR tasks for better results.

TABLE V. Comparison on the number of parameters of different models.

Model	**# Parameters**
UEE (Liu *et al.*, 2018)	22.42M
MTL-MBE (Hou *et al.*, 2020)	6.82M
EP-WUN (Lin *et al.*, 2023)	4.58M
WSRGlow (Zhang *et al.*, 2021)	229M
NU-Wave 2 (Han and Lee, 2022)	**1.70M**
VoiceFixer (Liu *et al.*, 2022)	122.07M
AERO (Mandel *et al.*, 2023)	19.43M
AudioSR (Liu *et al.*, 2024)	258.20M
Ours	25.04M

For the noise-robust SSR, when the sampling rates of source are flexible, we retrained VoiceFixer, a general speech restoration model, as our baseline model since current comparable models only support 8 kHz input signals. As shown in the following section of Table

4, in a comparison with VoiceFixer, our method outperforms it in all metrics for both 8 kHz to 16 kHz and 4-16 kHz to 16 kHz noise-robust BWE tasks. VoiceFixer aims to repair many distortions such as clipping, reverberation, and we find the speeches produced mismatch with the reference signal in terms of loudness, etc., which causes the degradation of its performance in objective metrics, but in subjective metrics, the scores of these speeches are still very high, which shows its repair is still very effective. We also summarize the number of parameters in each baseline model, and the results are in Table V. It is observed that our proposed model results in an increase in parameters, but this is acceptable due to the significant performance gain achieved.

TABLE VI. Test results of denoise-only task on DNS no-reverb noisy test set sampling at 16 kHz.

Method	**PESQ-NB**	**PESQ-WB**	**STOI(%)**
TSTNN (Wang *et al.*, 2021)	2.61	2.55	91.9
DPRNN (Luo *et al.*, 2020b)	2.68	2.57	92.5
TFT-Net (Tang *et al.*, 2020)	2.74	2.60	92.7
DCCRN (Hu *et al.*, 2020)	3.17	2.64	92.9
FullSubNet (Hao *et al.*, 2021)	3.28	2.72	95.3
DPT-FSNet (Dang *et al.*, 2022)	3.28	2.72	95.3
Ours	**3.29**	**2.80**	**96.0**

Additionally, we used our model to process DNS test set under wide-band environment at 16 kHz sampling rate, where the speech has the full bandwidth but contains noise in both the high and low frequency parts. To validate the facilitation of our joint optimization on a single task, we compare its performance with neural baseline models for only noise reduction. The results are depicted in Table VI. We observe that the proposed method improves both PESQ and STOI compared to the baseline models. Specifically, the narrow-band PESQ is slightly ahead of the best baseline model by 0.01, while the wide-band PESQ improves by 0.082, and the STOI achieves a performance of 96.0%, which is at least 0.7% higher than the baselines. Although the quality of the speech generated by our model is degraded due to the fact that it was not trained on 16 kHz noisy-clean data pairs compared to the noise-free and noise-robust SSR tasks, it still outperforms all the baselines. This indicates on the one hand that our model has high generalization capabilities and is able to repair unseen distortion types well, and on the other hand that our optimization for the joint task also benefits the single task.

C. Generalization Test

In order to better observe the generalization capability of the model, we tested the baseline models and proposed method using data from different source compared to training stage. In this case, for the noise-free case, we use the test set of the TIMIT (Garofolo *et al.*, 1993) and LibriTTS (Zen *et al.*, 2019) to perform 8 kHz to 16 kHz noise-free SSR task; and for the noise-robust case, we use the test set from Voicebank-DEMAND (Veaux *et al.*, 2013). It

is worth mentioning that all models involved in this comparison have not been trained with the data from these two datasets.

Tables VII, VIII, and IX show the test results for the noise-free case and the noise-robust case, respectively. Our model presents better performance on wide-band, with PESQ-WB significantly higher than the other baselines, and maintains the lead in other metrics as well. This indicates that our data augmentation approach allows the model to show a better

TABLE VII. Generalization test results on TIMIT.

Method	**PESQ-NB**	**PESQ-WB**	**STOI(%)**	**CSIG**	**CBAK**	**COVL**	**LSD**
WSRGlow (Zhang *et al.*, 2021)	4.087	2.180	98.5	3.558	3.425	2.916	1.146
NU-Wave2 (Han and Lee, 2022)	4.479	2.327	97.5	3.705	2.122	3.070	2.110
VoiceFixer (Liu *et al.*, 2022)	2.890	1.884	88.5	2.965	1.753	2.375	1.190
AudioSR (Liu *et al.*, 2024)	**4.491**	2.939	99.3	3.904	2.607	3.480	1.430
AERO (Mandel *et al.*, 2023)	4.481	3.401	**99.7**	4.226	4.261	3.870	1.176
Ours	4.489	**4.029**	**99.7**	**4.228**	**4.644**	**4.188**	**1.137**

TABLE VIII. Generalization test results on LibriTTS.

Method	PESQ-NB	PESQ-WB	STOI(%)	CSIG	CBAK	COVL	LSD
WSRGlow (Zhang *et al.*, 2021)	4.051	2.694	98.1	3.914	4.142	3.359	1.039
NU-Wave2 (Han and Lee, 2022)	4.237	2.682	94.1	3.108	2.811	2.932	1.391
VoiceFixer (Liu *et al.*, 2022)	3.194	2.773	93.9	3.186	2.813	2.890	1.137
AudioSR (Liu *et al.*, 2024)	4.293	2.728	98.6	3.774	3.533	3.308	1.113
AERO (Mandel *et al.*, 2023)	4.308	3.500	**99.4**	4.386	4.656	4.005	**0.988**
Ours	**4.377**	**3.647**	**99.4**	**4.412**	**4.752**	**4.118**	1.085

generalization performance for speech features from other channels, and this performance gain is observed in both the noisy and noise-free environments.

D. Performance on Compressed Speeches

Speech signal compression in real-world conditions involves reducing the data required to represent speech, which mainly include:

- Bit Compression: Reducing the depth of bit, leading to a loss of detail and fidelity.

TABLE IX. Generalization test results on Voiceband-DEMAND.

Method	PESQ-NB	PESQ-WB	STOI(%)	CSIG	CBAK	COVL	LSD
VoiceFixer (Liu *et al.*, 2022)	3.062	2.369	88.6	3.432	2.327	2.901	1.081
I-DTLN+AFiLM (Chen *et al.*, 2022)	3.059	2.090	89.3	1.877	**2.827**	1.925	1.460
Ours	**3.295**	**2.380**	**93.4**	**3.526**	2.356	**2.915**	**1.003**

- Sampling Rate Reduction: Lowering the sampling rate, which reduces audio resolution and high-frequency details.

- Data Compression Algorithms: Applying lossy compression techniques that remove parts of the audio signal deemed less perceptually important, often introducing artifacts (e.g., MP3, AAC).

These methods are crucial for efficient storage and transmission, especially in environments with limited bandwidth. However, they can affect the clarity and naturalness of the compressed speech.

The main objective of a SSR model is specifically designed to address the second type of compression, where the objective is to convert low sampling rate audio into high sampling rate audio, which means the tasks like transforming low bit depth speech clip to the high one or restoring the lossless speech from its lossy compression version are out of its capacity.

However, our model can be robust under these conditions involving compression, namely our system supports to deal with the speech with lower bit depth and lossy encoding format, but the system only predicts its lossy band, keeping the bit depth and format the same.

TABLE X. Performance valuation on compressed speech clips and downstream task.

Input	**Format**	**Bit Depth**	**Sampling Rate**	**PESQ-NB**	**PESQ-WB**	**STOI(%)**	**LSD**	**WAcc(%)**
Noisy	Lossless (.wav/flac)	8bit	8 kHz	3.103	1.981	87.9	1.004	——
Predict			16 kHz	3.253	2.192	89.1	0.820	——
Noisy		16bit	8 kHz	2.879	1.910	92.0	2.721	90.90
Predict			16 kHz	3.295	2.369	93.4	1.003	92.62
Reference				——	——	——	——	95.98
Noisy	Lossy (.mp3)	16bit	8 kHz	2.975	1.939	90.9	2.790	——
Predict			16 kHz	3.296	2.274	92.4	1.476	——

We have normalised the test set of Voicebank-DEMAND (Veaux *et al.*, 2013) to a lower bit depth (8bit) and a lossy compression format (MP3) respectively and the test results are in Table X. In this test, we upsample the 8 kHz signal to 16 kHz to calculate PESQ-WB and LSD. Table X shows that the performance of the model is not affected and it still significantly improves the signal quality.

E. Downstream Task Evaluation

In order to assess the effectiveness of proposed model in enhancing the performance of downstream tasks. By taking Automatic Speech Recognition (ASR) as an instance, we evaluate the ASR performance on original low-resolution, enhanced and a reference speech clips of Voicebank-DEMAND (Veaux *et al.*, 2013) test set, where we use the base version of Whisper [2] (Radford *et al.*, 2023) as the pre-trained ASR system in all cases. The results are also provided in Table X.

The experiment concludes that our method enhances the performance of ASR compared to the original lossy speech. These results demonstrate the model's potential to improve ASR robustness and reliability, confirming its value as a pre-processing step in real-world speech processing applications.

F. Ablation Studies

We conduct the ablation studies using the DNS no-reverb test set and 8 kHz to 16 kHz noise-robust SSR task, and the experiments are set to verify the influences of network components, loss functions, FFT bins, and resampling algorithm to the final performance.

The results are listed in Table XI. From the network structure point of view, when the gated convolution ('w/o GConv' in the table) is replaced by the general convolution, the network performance degrades due to the lack of 6-8 kHz details. If the LB is removed, the performance also decreases due to not utilising the time dimension information in the spectrogram tensor. When both of these changes work together, the accuracy of the model drops even more. The network achieves the best LSD performance when using only the MSTFT loss, but the PESQ is not as good as the optimal setting for either narrow-band or

TABLE XI. Results of ablation studies.

Method	**PESQ-NB**	**PESQ-WB**	**LSD**
w/o LBs	3.442	2.633	1.256
w/o GConv	3.445	2.630	1.262
w/o LBs and GConv	3.372	2.538	1.293
w/o adversarial training	3.453	2.658	**1.200**
w/o adversarial loss	3.313	2.484	1.242
w/o feature loss	2.941	1.840	1.309
FFT bins=128	3.274	2.459	1.272
FFT bins=256	3.238	2.369	1.301
w/o Algorithm 1	3.341	2.483	1.240
original settings	**3.554**	**2.777**	1.218

wide-band, being lower by 0.101 and 0.119, respectively. On top of this, the introduction of either the feature loss or the adversarial loss alone deteriorates the performance and moves the model even further away from the optimal performance. Also, if the number of FFT bins during the STFT operation is chosen larger or a direct downsampling function is used to produce the training data, the performance of the network is degraded as the input features become coarser. Therefore, the results of the ablation experiments illustrate that our proposed modules, adversarial training policy, and data augmentation approach improve the overall performance of the model on the test set, and also shows that the network performs best with FFT bins of 512, which is exactly the setting we used.

G. Spectrogram Comparison

Figure 8 - 10 are the comparisons of the spectrograms that are generated by different models on different tasks. On noise-free SSR task (see Figure 8), the result of our method is closer to the ground truth and presents no artifacts at the 4 kHz band, while other methods produce some bias at high-frequency part and has the unnatural transition band.

The similar situation also exists in the noise-robust SSR task (see Figure 9). Compared to our method, I-DTLN + AFiLM model (Figure 9(b)) only predicts a small part of the whole high frequency band and the VoiceFixer (Figure 9(c)) generates a spectrogram with a larger amplitude than the ground truth, causing the deviation. For 16 kHz to 16 kHz denoise task (Figure 10), baselines' results still produce residual noises in either low- or high-frequency parts, while our model generates a better result.

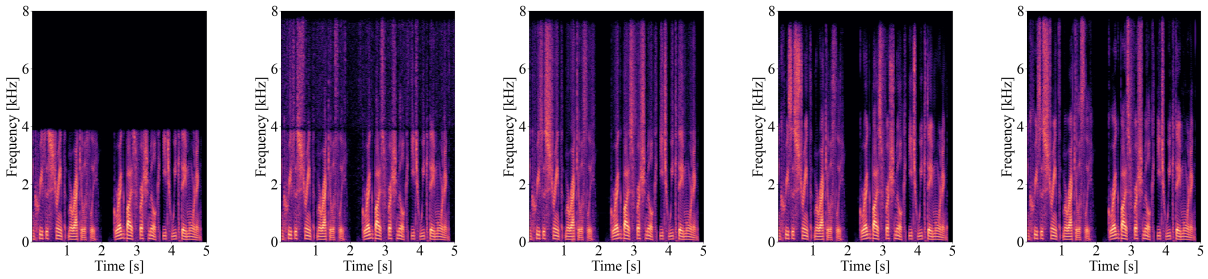

FIG. 8. (color online) Spectrograms of noise-free SSR task results. (a) Input; (b) Nu-Wave 2; (c) WSRGlow; (d) our method; (e) ground truth.

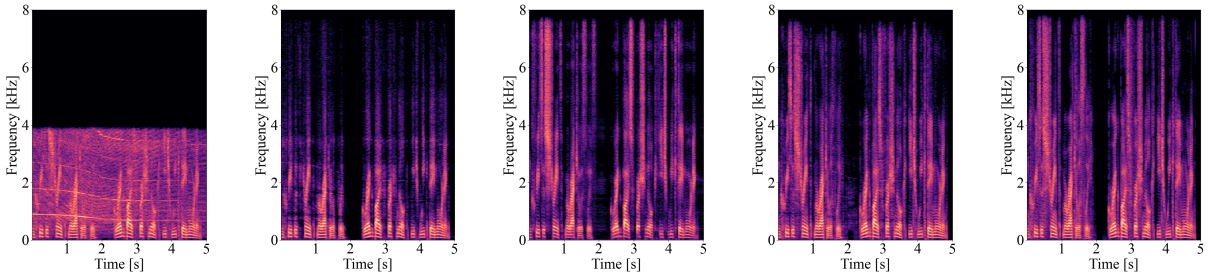

FIG. 9. (color online) Spectrograms of noise-robust SSR task results. (a) Input; (b) I-DTLN + AFiLM; (c) VoiceFixer; (d) our method; (e) ground truth.

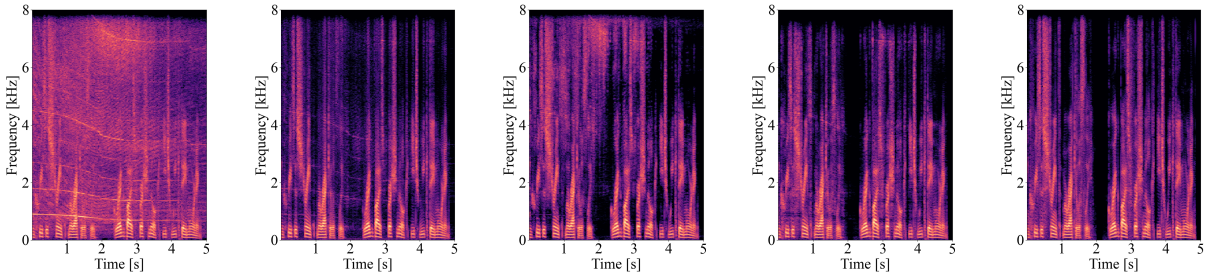

FIG. 10. (color online) Spectrograms of 16 kHz to 16 kHz denoise results. (a) Input; (b) DPRNN; (c) DCCRN; (d) our method; (e) ground truth.

VI. CONCLUSION

This paper proposes a novel noise-robust speech super-resolution model, termed SDNet. We introduce a U-shaped neural architecture generator, employing FTB, gated convolution, lattice blocks, and other modules, some of which are employed in the SSR field for the first time. Adversarial training is achieved through multi-scale discriminators with multiple loss functions, building robust reconstruction capability for the generator, augmented by a specialised data augmentation algorithm. The proposed model demonstrates superior performance in noise-free SSR, noise-robust SSR, and denoise-only tasks, for both fixed and flexible input sampling rates. Ablation studies demonstrate the effectiveness of our design choices. However, when training the model at higher resolutions such as 48 kHz, achieving denoising and SSR simultaneously becomes challenging, a common issue encountered by many models. Furthermore, the model's parameter count (25.04M) remains substantial. Future work will focus on lightweight, high resolution SSR, and considering the inclusion of music and other personalised datasets.

VII. AUTHOR DECLARATIONS

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

VIII. DATA AVAILABILITY

Data are available on request to the authors.

[1]https://sdnetdemo.github.io/

[2]https://github.com/openai/whisper

835, I.-T. P. (**2003**). "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," ITU-T Recommendation .

Bauer, P., and Fingscheidt, T. (**2008**). "An hmm-based artificial bandwidth extension evaluated by cross-language training and test," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, pp. 4589–4592.

Birnbaum, S., Kuleshov, V., Enam, Z., Koh, P. W. W., and Ermon, S. (**2019**). "Temporal film: Capturing long-range sequence dependencies with feature-wise modulations.," Advances in Neural Information Processing Systems **32**.

Chen, C.-W., Wang, W.-C., Ou, Y.-Y., and Wang, J.-F. (**2022**). "Deep learning audio super resolution and noise cancellation system for low sampling rate noise environment," in *2022 10th International Conference on Orange Technology (ICOT)*, IEEE, pp. 1–5.

Cheng, Y. M., O'Shaughnessy, D., and Mermelstein, P. (**1994**). "Statistical recovery of wideband speech from narrowband speech," IEEE Transactions on Speech and Audio Processing **2**(4), 544–548.

Dang, F., Chen, H., and Zhang, P. (**2022**). "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)*, pp. 6857–6861, doi: 10.1109/ICASSP43922.2022.9746171.

Dubey, H., Aazami, A., Gopal, V., Naderi, B., Braun, S., Cutler, R., Ju, A., Zohourian, M., Tang, M., Golestaneh, M. *et al.* (**2024**). "Icassp 2023 deep noise suppression challenge," IEEE Open Journal of Signal Processing .

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., and Pallett, D. S. (**1993**). "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," NASA STI/Recon technical report n **93**, 27403.

Han, S., and Lee, J. (**2022**). "NU-Wave 2: A General Neural Audio Upsampling Model for Various Sampling Rates," in *Proc. Interspeech 2022*, pp. 4401–4405, doi: 10.21437/Interspeech.2022-45.

Hao, X., Su, X., Horaud, R., and Li, X. (**2021**). "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6633–6637, doi: 10.1109/ICASSP39728.2021.9414177.

Haws, D., and Cui, X. (**2019**). "Cyclegan bandwidth extension acoustic modeling for automatic speech recognition," in *Proc. ICASSP*, pp. 6780–6784, doi: 10.1109/ICASSP.2019.8682760.

Hernandez-Olivan, C., Saito, K., Murata, N., Lai, C.-H., Martínez-Ramírez, M. A., Liao, W.-H., and Mitsufuji, Y. (**2024**). "Vrdmg: Vocal restoration via diffusion posterior sampling with multiple guidance," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 596–600, doi: 10.1109/ICASSP48485.2024.10446423.

Hou, N., Xu, C., Zhou, J. T., Chng, E. S., and Li, H. (**2020**). "Multi-Task Learning for End-to-End Noise-Robust Bandwidth Extension," in *Proc. Interspeech 2020*, pp. 4069–4073, doi: 10.21437/Interspeech.2020-2022.

Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., Wu, J., Zhang, B., and Xie, L. (**2020**). "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in *Proc. Interspeech 2020*, pp. 2472–2476, doi: 10.21437/Interspeech.2020-2537.

Hu, Y., and Loizou, P. C. (**2007**). "Evaluation of objective quality measures for speech enhancement," IEEE Transactions on audio, speech, and language processing **16**(1), 229–238.

Kim, S.-B., Lee, S.-H., Choi, H.-Y., and Lee, S.-W. (**2024**). "Audio super-resolution with robust speech representation learning of masked autoencoder," IEEE/ACM Transactions on Audio, Speech, and Language Processing **32**, 1012–1022, doi: 10.1109/TASLP.2023.3349053.

Kong, J., Kim, J., and Bae, J. (**2020**). "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," Advances in neural information processing systems **33**, 17022–17033.

Kuleshov, V., Enam, S. Z., and Ermon, S. (**2017**). "Audio super resolution using neural networks," arXiv preprint arXiv:1708.00853 .

Kumar, K., Kumar, R., De Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., De Brebisson, A., Bengio, Y., and Courville, A. C. (**2019**). "Melgan: Generative adversarial networks for conditional waveform synthesis," Advances in neural information processing systems **32**.

Li, K., and Lee, C.-H. (**2015**). "A deep neural network approach to speech bandwidth expansion," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 4395–4399.

Li, Y., Tagliasacchi, M., Rybakov, O., Ungureanu, V., and Roblek, D. (**2021**). "Real-time speech frequency bandwidth extension," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 691–695.

Lim, T. Y., Yeh, R. A., Xu, Y., Do, M. N., and Hasegawa-Johnson, M. (**2018**). "Time-frequency networks for audio super-resolution," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 646–650.

Lin, Y.-T., Su, B.-H., Lin, C.-H., Kuo, S.-C., Jang, J.-S. R., and Lee, C.-C. (**2023**). "Noise-Robust Bandwidth Expansion for 8K Speech Recordings," in *Proc. INTERSPEECH 2023*, pp. 5107–5111, doi: 10.21437/Interspeech.2023-857.

Ling, Z.-H., Ai, Y., Gu, Y., and Dai, L.-R. (**2018**). "Waveform modeling and generation using hierarchical recurrent neural networks for speech bandwidth extension," IEEE/ACM Transactions on Audio, Speech, and Language Processing **26**(5), 883–894, doi: 10.1109/TASLP.2018.2798811.

Liu, B., Tao, J., and Zheng, Y. (**2018**). "A novel unified framework for speech enhancement and bandwidth extension based on jointly trained neural networks," in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, IEEE, pp. 11–15.

Liu, H., Chen, K., Tian, Q., Wang, W., and Plumbley, M. D. (**2024**). "Audiosr: Versatile audio super-resolution at scale," in *ICASSP 2024 - 2024 IEEE International Conference*

654 *on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1076–1080, doi: 10.1109/
655 ICASSP48485.2024.10447246.

656 Liu, H., Liu, X., Kong, Q., Tian, Q., Zhao, Y., Wang, D., Huang, C., and Wang, Y.
657 (**2022**). "VoiceFixer: A Unified Framework for High-Fidelity Speech Restoration," in *Proc.*
658 *Interspeech 2022*, pp. 4232–4236, doi: 10.21437/Interspeech.2022-11026.

659 Luo, X., Qu, Y., Xie, Y., Zhang, Y., Li, C., and Fu, Y. (**2022**). "Lattice network for
660 lightweight image restoration," IEEE Transactions on Pattern Analysis and Machine In-
661 telligence **45**(4), 4826–4842.

662 Luo, X., Xie, Y., Zhang, Y., Qu, Y., Li, C., and Fu, Y. (**2020**a). "Latticenet: Towards
663 lightweight image super-resolution with lattice block," in *Computer Vision–ECCV 2020:*
664 *16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*,
665 Springer, pp. 272–289.

666 Luo, Y., Chen, Z., and Yoshioka, T. (**2020**b). "Dual-path rnn: Efficient long sequence
667 modeling for time-domain single-channel speech separation," in *ICASSP 2020 - 2020 IEEE*
668 *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 46–50,
669 doi: 10.1109/ICASSP40776.2020.9054266.

670 Mandel, M., Tal, O., and Adi, Y. (**2023**). "Aero: Audio super resolution in the spectral
671 domain," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and*
672 *Signal Processing (ICASSP)*, IEEE, pp. 1–5.

673 Moliner, E., Lehtinen, J., and Välimäki, V. (**2023**). "Solving audio inverse problems with
674 a diffusion model," in *ICASSP 2023 - 2023 IEEE International Conference on Acous-*
675 *tics, Speech and Signal Processing (ICASSP)*, pp. 1–5, doi: 10.1109/ICASSP49357.2023.

10095637.

Moliner, E., and Välimäki, V. (**2023**). "Behm-gan: Bandwidth extension of historical music using generative adversarial networks," IEEE/ACM Transactions on Audio, Speech, and Language Processing **31**, 943–956, doi: 10.1109/TASLP.2022.3190726.

Moreno, P., Raj, B., and Stern, R. (**1996**). "A vector taylor series approach for environment-independent speech recognition," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, Vol. 2, pp. 733–736 vol. 2, doi: 10.1109/ICASSP.1996.543225.

Nguyen, V.-A., Nguyen, A. H., and Khong, A. W. (**2022**). "Tunet: A block-online bandwidth extension model based on transformers and self-supervised pretraining," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 161–165.

Nour-Eldin, A. H., and Kabal, P. (**2009**). "Combining frontend-based memory with mfcc features for bandwidth extension of narrowband speech," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, pp. 4001–4004.

Park, K.-Y., and Kim, H. S. (**2000**). "Narrowband to wideband conversion of speech using gmm based transformation," in *2000 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 00CH37100)*, IEEE, Vol. 3, pp. 1843–1846.

Pulakka, H. *et al.* (**2013**). "Development and evaluation of artificial bandwidth extension methods for narrowband telephone speech," .

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (**2023**). "Robust speech recognition via large-scale weak supervision," in *International conference*

on machine learning, PMLR, pp. 28492–28518.

Rakotonirina, N. C. (**2021**). "Self-attention for audio super-resolution," in *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*, IEEE, pp. 1–6.

Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. (**2001**). "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, IEEE, Vol. 2, pp. 749–752.

Seltzer, M. L., Acero, A., and Droppo, J. (**2005**). "Robust bandwidth extension of noise-corrupted narrowband speech," in *Proc. Interspeech 2005*, pp. 1509–1512, doi: 10.21437/Interspeech.2005-529.

Seo, H., Kang, H.-G., and Soong, F. (**2014**). "A maximum a posterior-based reconstruction approach to speech bandwidth expansion in noise," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6087–6091, doi: 10.1109/ICASSP.2014.6854773.

Shuai, C., Shi, C., Gan, L., and Liu, H. (**2023**). "mdctGAN: Taming transformer-based GAN for speech super-resolution with Modified DCT spectra," in *Proc. INTERSPEECH 2023*, pp. 5112–5116, doi: 10.21437/Interspeech.2023-113.

Stoller, D., Ewert, S., and Dixon, S. (**2018**). "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," in *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018*, edited by E. Gómez, X. H. 0001, E. Humphrey, and E. Benetos, pp. 334–340,

http://ismir2018.ircam.fr/doc/pdfs/205_Paper.pdf.

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (**2011**). "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," IEEE Transactions on Audio, Speech, and Language Processing **19**(7), 2125–2136, doi: 10.1109/TASL.2011.2114881.

Taher, T., Mamun, N., and Hossain, M. A. (**2023**). "A joint bandwidth expansion and speech enhancement approach using deep neural network," in *2023 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, IEEE, pp. 1–4.

Tang, C., Luo, C., Zhao, Z., Xie, W., and Zeng, W. (**2020**). "Joint time-frequency and time domain learning for speech enhancement," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, edited by C. Bessiere, International Joint Conferences on Artificial Intelligence Organization, pp. 3816–3822, https://doi.org/10.24963/ijcai.2020/528, doi: 10.24963/ijcai.2020/528, main track.

Taylor, A. M., and Reby, D. (**2010**). "The contribution of source–filter theory to mammal vocal communication research," Journal of Zoology **280**(3), 221–236.

Valentini-Botinhao, C., Wang, X., Takaki, S., and Yamagishi, J. (**2016**). "Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System Using Deep Recurrent Neural Networks," in *Proc. Interspeech 2016*, pp. 352–356, doi: 10.21437/Interspeech.2016-159.

Veaux, C., Yamagishi, J., and King, S. (**2013**). "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *2013 International Con-*

ference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), pp. 1–4, doi: 10.1109/ICSDA.2013.6709856.

Wang, H., and Wang, D. (**2021**). "Towards robust speech super-resolution," IEEE/ACM transactions on audio, speech, and language processing **29**, 2058–2066.

Wang, K., He, B., and Zhu, W.-P. (**2021**). "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)*, pp. 7098–7102, doi: 10.1109/ICASSP39728.2021.9413740.

Xu, C., Tan, G., and Ying, D. (**2023**). "Time-frequency network combining batch attention and spatial attention for speech bandwidth extension," Applied Acoustics **211**, 109582, https://www.sciencedirect.com/science/article/pii/S0003682X23003808, doi: https://doi.org/10.1016/j.apacoust.2023.109582.

Yin, D., Luo, C., Xiong, Z., and Zeng, W. (**2020**). "Phasen: A phase-and-harmonics-aware speech enhancement network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, pp. 9458–9465.

Yoneyama, R., Yamamoto, R., and Tachibana, K. (**2023**). "Nonparallel high-quality audio super resolution with domain adaptation and resampling cyclegans," in *Proc. ICASSP*, pp. 1–5.

Yu, C.-Y., Yeh, S.-L., Fazekas, G., and Tang, H. (**2023**). "Conditioning and sampling in variational diffusion models for speech super-resolution," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp.

1–5.

Yu, G., Zheng, X., Li, N., Han, R., Zheng, C., Zhang, C., Zhou, C., Huang, Q., and Yu, B. (**2024**). "Bae-net: a low complexity and high fidelity bandwidth-adaptive neural network for speech super-resolution," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 571–575, doi: 10.1109/ICASSP48485.2024.10446439.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. S. (**2019**). "Free-form image inpainting with gated convolution," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4471–4480.

Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., Chen, Z., and Wu, Y. (**2019**). "Libritts: A corpus derived from librispeech for text-to-speech," in *Interspeech 2019*, pp. 1526–1530, doi: 10.21437/Interspeech.2019-2441.

Zhang, K., Ren, Y., Xu, C., and Zhao, Z. (**2021**). "WSRGlow: A Glow-Based Waveform Generative Model for Audio Super-Resolution," in *Proc. Interspeech 2021*, pp. 1649–1653, doi: 10.21437/Interspeech.2021-892.

Zhao, L., Zhu, W., Li, S., Luo, H., Zhang, X.-L., and Rahardja, S. (**2024**). "Multi-resolution convolutional residual neural networks for monaural speech dereverberation," IEEE/ACM Transactions on Audio, Speech, and Language Processing **32**, 2338–2351, doi: 10.1109/TASLP.2024.3385270.