This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/TASLP.2024.3468026, IEEE/ACM Transactions on Audio Speech and Language Processing

Cross Domain Optimization for Speech Enhancement: Parallel or Cascade?

Liang Wan, Hongqing Liu, *Senior Member, IEEE,* Liming Shi, *Member, IEEE,* Yi Zhou, *Member, IEEE,* and Lu Gan, *Senior Member, IEEE*

***Abstract*—This paper introduces five novel deep-learning architectures for speech enhancement. Existing methods typically use time-domain, time-frequency representations, or a hybrid approach. Recognizing the unique contributions of each domain to feature extraction and model design, this study investigates the integration of waveform and complex spectrogram models through cross-domain fusion to enhance speech feature learning and noise reduction, thereby improving speech quality. We examine both cascading and parallel configurations of waveform and complex spectrogram models to assess their effectiveness in speech enhancement. Additionally, we employ an orthogonal projection-based error decomposition technique and manage the inputs of individual sub-models to analyze factors affecting speech quality. The network is trained by optimizing three specific loss functions applied across all sub-models. Our experiments, using the DNS Challenge (ICASSP 2021) dataset, reveal that the proposed models surpass existing benchmarks in speech enhancement, offering superior speech quality and intelligibility. These results highlight the efficacy of our cross-domain fusion strategy. We provide a demo page containing enhanced audio clips from different models at https://wanliangdaxia.github.io/.**

***Index Terms*—speech enhancement, waveform, time-frequency, complex domain, cross-domain speech.**

I. Introduction

SPEECH enhancement aims to improve the quality of targeted speech signals, with broad applications including teleconferencing, hearing aids, and other types of communications. Traditional statistical methods apply gains or filters to noisy signals [1], but often yield suboptimal performance. Deep learning has revolutionized audio processing, giving rise to innovative speech enhancement methods [2]–[5] based on data-driven supervised learning [6]. These approaches effectively reduce noise, especially for non-stationary noise, gaining significant attention.

Deep neural network (DNN)-based speech enhancement methods are broadly divided into time-domain [7]–[9] and time-frequency (TF) [6], [10], [11] approaches. Time domain strategies offer end-to-end training to predict clean speech waveforms, without explicitly using short time Fourier transform (STFT) as in TF domain methods. It is important to acknowledge that discrete cosine transform (DCT) and wavelet transforms can also be considered part of TF domain methods. However, STFT is selected here due to its widespread usage and established presence. Nevertheless, this approach might not fully harness the auditory patterns identified in TF spectrograms, potentially compromising the overall quality of speech enhancement.

On the other hand, TF domain approaches utilize two-dimensional spectrograms, enhancing the differentiation between speech and noise. These methods include masking-based and mapping-based strategies. Masking-based techniques, such as the ideal binary mask (IBM) [12], ideal ratio mask (IRM) [13], and spectral magnitude mask (SMM) [13], focus on the magnitude discrepancies between clean and noisy speech, often overlooking phase information. To address this, the complex ratio mask (CRM) [14] aims to refine both the real and imaginary components of the spectrogram, facilitating a more accurate reconstruction of the speech signal. Mapping-based speech enhancement [15], meanwhile, directly transforms noisy speech spectrograms into their clean counterparts, leveraging the full spectrum of information, including phase, to achieve a higher fidelity in speech signal restoration. This holistic approach enables a nuanced enhancement, capturing the intricate details necessary for high-quality speech reconstruction.

Drawing on the curriculum learning concept [16], multi-stage mapping algorithms have emerged as a significant concept in speech front-end tasks. These algorithms divide the original mapping challenge into several simpler subtasks, enabling a step-by-step enhancement in performance. This progressive approach leads to superior solutions via an iterative refinement process. Tzinis et al. [17] introduced a two-step architectural framework for source separation. Their approach deviates from direct source separation, focusing initially on learning a latent representation of speech. Subsequently, the separation task is executed within this acquired latent embedding space, thereby refining the process. The multi-stage self-attentive temporal convolutional network (SA-TCN) [18] method employs a multi-stage learning paradigm for speech enhancement, utilizing a layered structure. Each layer comprises a self-attention block succeeded by a series of TCN blocks with progressively increasing dilation factors. At every stage, a refined prediction is generated, which is further polished in subsequent stages. To ensure the retention

Liang Wan was with School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, and he now is with China Telecom Corporation Limited Chongqing Branch, E-mail: wanliang1996@gmail.com.

Hongqing Liu is with Chongqing Key Lab of Mobile Communications Technology and Intelligent Speech and Audio Research Lab, Chongqing University of Posts and Telecommunications, Chongqing 400065, China. E-mail: hongqingliu@outlook.com. (*Corresponding author*)

Liming Shi and Yi Zhou are with School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China. E-mail: shilm@cqupt.edu.cn; zhouy@cqupt.edu.cn.

Lu Gan is with College of Engineering, Design and Physical Science, Brunel University, London UB8 3PH, U.K. E-mail: lu.gan@brunel.ac.uk.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/TASLP.2024.3468026, IEEE/ACM Transactions on Audio Speech and Language Processing

of original data, a feature fusion block is incorporated at the onset of later stages. Zhang et al. [19] proposed a dual-stage framework in speech enhancement integrating multiple training targets. The initial stage involves dual-branch training, where one branch is dedicated to predicting the complex spectrum and the other to the IRM. The next stage leverages the enhanced magnitude from the first stage for prior Signal-to-Noise Ratio (SNR) prediction. In a novel approach, Wang and Wang [20] introduced a neural cascade architecture (NCA) that capitalizes on the benefits of cross-domain speech representations. The NCA consists of three distinct modules that process the spectral magnitude, waveform, and complex spectrogram individually. Each module not only processes the output of its predecessor but also references the original noisy input, ensuring a comprehensive and nuanced enhancement process.

We contend that each domain, whether time or TF, offers distinct information, where the time domain may capture richer harmonic characteristics essential for speech structure while the TF domain offers superior frequency resolution for better speech-noise discrimination. Leveraging this complementary information might enhance final speech enhancement performance. Pursuing this line of thought, our work introduces innovative strategies to fuse information from different domains within both cascade and parallel architectures. This exploration aims to ascertain which structural configuration most effectively enhances speech quality. Concurrently, we incorporate both waveform and complex spectrogram into our proposed model to capitalize on their unique learning capacities for diverse features. This combination significantly improves the overall learning prowess of our model, enabling a more comprehensive and nuanced approach to speech enhancement.

The contributions of this work are summarized as follows:

- **Innovative Performance Analysis:** We conducted the first in-depth performance analysis of cascaded and parallel architectures specifically designed for speech enhancement. Beyond providing valuable insights into the feature propagation mechanisms of these architectures, our study introduces a novel analytical framework that better leverages the learning capabilities of different domain models. This understanding allowed us to better leverage the strengths of different domain models, ultimately achieving superior speech quality and advancing the field significantly.
- **New Insights into Architectures:** Our research offers a detailed investigation into the role of outputs from various tiers of cascaded architectures and their influence on subsequent modules' learning efficiency. By introducing a controlled information flow mechanism—where each level strategically incorporates or disregards its predecessor's output—we provide a new, visual understanding of how feature transmission affects model performance. This analysis not only deepens our comprehension of systems but also proposes a novel method for optimizing these architectures for better performance in speech enhancement.
- **Feature Integration:** We developed a novel structure that effectively integrates distinct feature representations from different domains—highlighting the harmonic richness of time domain features and the superior speech-noise discrimination of TF domain features. This innovative fusion of complementary domain strengths creates a synergistic effect, significantly enhancing both noise suppression and speech preservation capabilities. Our approach demonstrates a new way of capitalizing on the unique advantages of each domain, leading to a superior performance in the speech enhancement task.

The rest of the paper is structured as follows: Section II presents the related algorithms, which cover single domain and multiple domain structures. In Section III, we describe our concept, including time domain and TF domain modules, and how we leverage different domain information, leading to the cascaded and parallel structures. Section IV presents the experimental results and associated discussions on the performance of the proposed network compared with the baseline and different models in various scenarios. Finally, we conclude and suggest topics for future research in Section V.

II. RELATED WORK

A. Single Domain Network Architectures

Conv-Tasnet [7], renowned for its proficiency in time-domain processing, employs 1-D convolutional layers to encode waveform inputs. This encoding process creates efficient representations that are pivotal for precise speech estimation. Subsequently, the encoded data undergoes decoding through transposed convolutional layers, reconstructing the original waveform. However, processing exceptionally long sequences in the time domain presents challenges. To address this, deeper convolutional architectures like Wave-U-Net [9] are employed, which can compress features effectively. The Wave-U-Net architecture synergizes elements from both Conv-Tasnet and the U-Net architecture. Drawing inspiration from the spectrogram-based U-Net [21], [22] methodology, Wave-U-Net uses a series of downsampling and upsampling blocks to form predictions. A notable aspect of this architecture is the halving of time resolution at each network level, which is instrumental in the model's enhanced capability to improve speech quality. This systematic reduction in time resolution at each stage plays a crucial role in distilling and refining the speech signal, thereby augmenting the overall speech enhancement process.

Speech enhancement has witnessed remarkable advancements owing to carefully designed network architectures. A notable shift in recent developments is the incorporation of phase information in TF domain networks. Within this spectrum, the convolution recurrent network (CRN) [23] has emerged as a prominent convolution encoder-decoder (CED) architecture, highly regarded for its efficacy in speech enhancement tasks.

Traditionally, speech enhancement techniques primarily relied on real spectrum inputs to estimate a real mask using neural networks. Tan et al. revolutionized this approach with the introduction of CSM [15], an innovative structure featuring one encoder and two decoders. This advanced architecture

facilitates the estimation of both real and imaginary components of the spectrum, significantly improving speech enhancement performance. However, this method treats real and imaginary parts as distinct input channels, utilizing a shared real-valued convolution filter for real-valued convolution operations, which does not comply with the rules of complex multiplication. Consequently, there is a tendency for networks to independently learn real and imaginary components without integrating prior knowledge effectively.

Addressing these limitations, the deep complex convolution recurrent network (DCCRN) [5] introduces substantial enhancements to the CRN model. DCCRN integrates complex CNN and complex batch normalization layers within both the encoder and decoder segments. Moreover, it contemplates substituting traditional long short-term memory (LSTM) with complex LSTM. The complex module in DCCRN adeptly models the correlation between magnitude and phase, simulating complex multiplication. This innovative approach facilitates a more seamless integration of real and imaginary components within the network structure. Notably, DCCRN demonstrated exceptional performance by achieving the highest mean opinion score (MOS) in the subjective listening test at the Interspeech 2020 deep noise suppression challenge [24], underscoring its proficiency in speech enhancement.

B. Multiple Domain Network Architectures

Recent studies have focused on exploring the combination of various speech representations to enhance speech enhancement performance. For example, some approaches aim to enhance speech through multiple stages, where each stage operates in a specific signal domain.

Hao *et al.* [25] introduced a novel two-stage speech enhancement technique that combines binary masking with spectrogram inpainting. Initially, a binary mask is created by applying a hard threshold to a soft mask, designed to isolate time-frequency points heavily influenced by intense noise. Subsequently, the spectrogram inpainting stage utilizes a CNN featuring partial convolution to refine the previously masked spectrogram. Li et al. [26] proposed a dual-stage network, where the initial stage is dedicated to estimating the magnitude spectrum. The subsequent stage then undertakes complex spectral mapping, leveraging both the predicted magnitude spectrum and the original noisy spectrum. This approach underscores the synergy between magnitude estimation and spectral mapping in speech enhancement.

In an effort to merge different representation domains within the loss function for enhanced results, Wang and Wang developed a NCA. Unlike other models that focus on one or two speech representation domains, the NCA incorporates three training targets in DNN-based speech enhancement. The NCA framework comprises three modules: CRN-Mask, UNet-Time, and CRN-Complex. Each module aligns with popular design strategies from contemporary speech enhancement research. The CRN-Mask module inputs magnitude features and predicts the IRM. The subsequent module, UNet-Time, processes two time-domain inputs, one from the inverse Short-Time Fourier Transform (iSTFT) of the masked spectrogram generated by CRN-Mask, and the other being the original noisy waveform. This design aims to reduce estimation errors and distortions from the preceding stage. The final module, CRN-Complex, takes the noisy complex spectrogram and the output of UNet-Time as inputs. All modules are optimized simultaneously using a triple-domain loss function. In addition, the NCA is trained end-to-end, simplifying the training process compared to other multi-stage models that often require complex training strategies like pretraining and fine-tuning. Experimental evaluations indicate that NCA significantly surpasses previous strong baselines in speech enhancement performance, demonstrating its effectiveness in integrating cross-domain speech representations.

C. Projection-based Decomposition

Iwamoto introduced an innovative analysis method for decomposing the estimation errors in speech enhancement by using orthogonal projections [27]. This technique, previously applied in the performance evaluation of speech enhancement and separation tasks [28], facilitates the partitioning of errors into two distinct components: the noise component (e_{noise}) and the artifact component (e_{artif}). These components are derived by projecting the estimation errors onto two subspaces: one is the speech-noise subspace formed by the speech and noise signals, and the other is a subspace orthogonal to the speech-noise subspace.

Denote y as the time-domain waveform of the observed signal. This signal is modeled as $y = s+n$, where s represents the source signal and n is the background noise. When the observed signal y is input, the enhanced signal $\hat{s}$ is estimated as $\hat{s} = SE(y)$, with $SE(\cdot)$ representing the speech enhancement module. The estimated signal $\hat{s}$ naturally includes estimation errors. Vincent et al. [28] suggested decomposing the estimated signal into three components, given by

$$\hat{s} = s_{target} + e_{noise} + e_{artif}, \tag{1}$$

where s_{target} is the target source component, projected on the signal subspace, and e_{noise} denotes the noise component, projected on the noise subspace, and e_{artif} represent artifact error component.

Iwamoto's innovative analysis technique employs orthogonal projection-based error decomposition to dissect and understand the impact of different error types on ASR performance. In this approach, the noise component e_{noise}, a mixture of speech and noise signals, represents naturally occurring sounds. These "natural" signals might have a limited effect on ASR performance since similar noise components are often present in training datasets. Conversely, the artifact component e_{artif} consists of signals that are not a linear combination of speech and noise. These "unnatural" signals, characterized by their diversity and rarity in training data, are believed to impart a sense of unnaturalness and significantly degrade perceived quality. Utilizing the orthogonal projection-based error decomposition, Iwamoto's analysis aims to validate hypotheses regarding the distinct impacts of these error types on ASR. He proposes manually manipulating the balance of noise and artifact errors in enhanced signals and conducting experiments to observe their respective effects on ASR performance.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/TASLP.2024.3468026, IEEE/ACM Transactions on Audio Speech and Language Processing

TABLE I
DETAILED PARAMETERS FOR TIME DOMAIN MODULE, WHERE THE HYPERPARAMETERS ARE ORDERED IN THE FASHION OF *Kernel Size, Stride, Number of Filters*.

layer name	input size	hyperparameters	output size
$encoder_1$	$(1*N)*16384$,	$15,1,24$,	$24*8192$
$encoder_2$	$24*8192$	$15,1,48$	$48*4096$
$encoder_3$	$48*4096$	$15,1,72$	$72*2048$
$encoder_4$	$72*2048$	$15,1,96$	$96*1024$
$encoder_5$	$96*1024$	$15,1,120$	$120*512$
$encoder_6$	$120*512$	$15,1,144$	$144*256$
$encoder_7$	$144*256$	$15,1,168$	$168*128$
$encoder_8$	$168*128$	$15,1,192$	$192*64$
$encoder_9$	$192*64$	$15,1,216$	$216*32$
$encoder_{10}$	$216*32$	$15,1,240$	$240*16$
$encoder_{11}$	$240*16$	$15,1,264$	$264*8$
$encoder_{12}$	$264*8$	$15,1,288$	$288*4$
$Convolution_1$	$288*4$	$15,1,288$	$288*4$
$decoder_1$	$576*8$	$5,1,288$	$288*8$
$decoder_2$	$552*16$	$5,1,264$	$264*16$
$decoder_3$	$504*32$	$5,1,240$	$240*32$
$decoder_4$	$456*64$	$5,1,216$	$216*64$
$decoder_5$	$408*128$	$5,1,192$	$192*128$
$decoder_6$	$360*256$	$5,1,168$	$168*256$
$decoder_7$	$312*512$	$5,1,144$	$144*512$
$decoder_8$	$264*1024$	$5,1,120$	$120*1024$
$decoder_9$	$216*2048$	$5,1,96$	$96*20248$
$decoder_{10}$	$168*4096$	$5,1,72$	$72*4096$
$decoder_{11}$	$120*8192$	$5,1,48$	$48*8192$
$decoder_{12}$	$72*16384$	$5,1,24$	$24*16384$
$Convolution_2$	$25*16384$	$1,1,1$	$1*16384$

Following Iwamoto's approach, our work also applies this orthogonal decomposition methodology to analyze how our proposed model affects speech quality. By varying the signal-to-noise ratio of the input, we aim to facilitate the model's ability to learn speech features and enhance speech quality. Additionally, we will examine the individual impacts of e_{artif} and e_{noise} on speech quality by adjusting their proportions. This will be achieved through the introduction of noisy signals, enabling a detailed investigation into how each type of error contributes to the overall performance of speech enhancement.

III. PROPOSED METHOD

As discussed earlier, each domain provides different information and that information need to be combined to improve the performance. In the next subsections, we first respectively introduce the time domain and TF domain modules used in this work, and then we show how we fuse different domain information either in cascade or parallel ways.

A. Time Domain Module

For the time domain module of our network, we employ a strategy that involves the use of downsampling (DS) blocks to extract high-level features at progressively coarser time scales. This is achieved by gradually increasing the number of DS blocks, which allows the network to capture a broad range of features from different time scales. These higher-level features are then integrated with local, high-resolution features that were computed at earlier stages. This integration is facilitated by upsampling (US) blocks, which serve to combine these diverse features effectively.

Our network is structured to include a total of 12 levels, with each subsequent level operating at half the time resolution of its predecessor. This hierarchical structure enables the network to process information at multiple scales, enhancing its ability to make accurate predictions. The DS blocks play a crucial role in this architecture by selectively discarding features at every other time step, effectively reducing the time resolution by half. Conversely, the US blocks work to increase the time resolution by a factor of two, employing linear interpolation to achieve upsampling in the time direction.

Additionally, our network incorporates Concat blocks, which are utilized to concatenate the high-level features processed at the current stage with more localized features. This concatenation is critical for preserving both the global and local characteristics of the input signal, ensuring a comprehensive feature representation. The specific parameters for the time domain module are outlined in Table I, where N denotes the number of channels in the input signal. This detailed parameter setup is designed to optimize the network's performance in processing and enhancing time-domain signals.

To train the network, we employ the mean square error (MSE) loss function to measure the estimation error from the signal, which is given by

$$L_{MSE} = MSE(S, \widetilde{S}), \tag{2}$$

where S represents the clean speech and $\widetilde{S}$ is the recovered one.

B. Complex TF Domain Module

The complex spectrogram module in our network is built on a primarily causal Convolutional Encoder-Decoder architecture, augmented with two unidirectional LSTM layers strategically positioned between the encoder and the decoder. The LSTM layers are specifically designed to capture and effectively model temporal dependencies within the data. This module processes waveform signals as inputs, which are initially transformed into their real and imaginary components using Conv-STFT. For the Conv-STFT operation, we utilize the Hanning window, setting the window length at 400 samples and the window shift at 100 samples. The real and imaginary parts obtained from this process are concatenated along the channel dimension, forming a comprehensive input for the encoder.

The encoder itself comprises six Conv2d blocks, tasked with extracting high-level features from the input and reducing its resolution. The output channels for each layer in the encoder are structured as $\{16,32,64,128,128,256\}$. We set the kernel size and stride of the Conv2d blocks to (5,2) and (2,1), respectively, and the LSTM units have 256 hidden layers. Subsequently, the decoder utilizes the low-resolution features processed by the encoder and reconstructs them to their original size, maintaining a symmetric structure between the encoder and decoder.

Each Conv2d block within the encoder/decoder is a sequence of a convolutional or deconvolutional layer, followed by batch normalization and an activation function. The architecture also incorporates skip-connections, which are pivotal

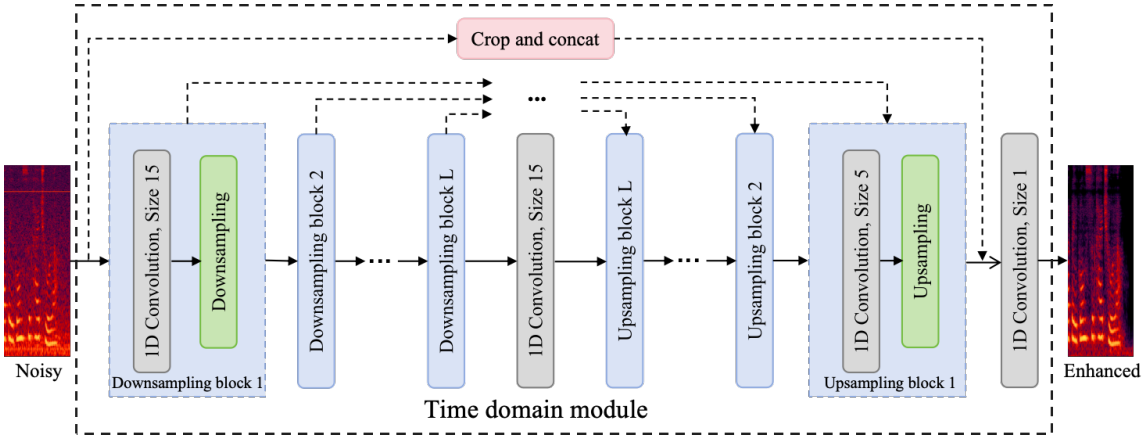


Fig. 1. **Time domain module**: Downsampling blocks compute higher-level features over coarser time scales. Upsampling blocks upsample temporally by a factor of two via linear interpolation. Concatenation blocks concatenate current high-level features with more local features

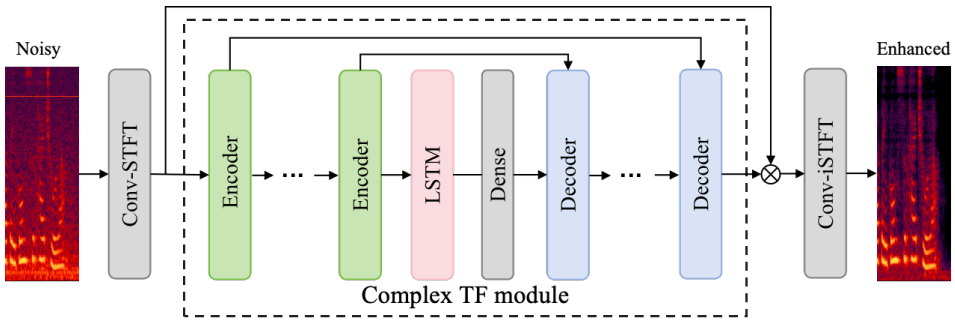


Fig. 2. **Complex TF module**: The input waveform is processed via Convolutional Short-Time Fourier Transform (Conv-STFT) [29] to obtain real and imaginary components. Convolutional and deconvolutional modules initialized with STFT kernels analyse and synthesise the waveforms before input to the network for loss computation. The encoder blocks extract high-level features and reduce resolution; the decoder blocks upsample the features to match the original input dimensions.

in enhancing gradient flow and establishing a linkage between the encoder and decoder components, thus facilitating more efficient learning and reconstruction.

The details of the parameter setup for this complex TF module are presented in Table II, where N denotes the number of channels in the input signal. This setup is carefully designed to optimize the processing and representation of complex spectrogram data, ensuring efficient feature extraction and reconstruction within the module.

During training, the complex spectrogram module estimates the CRM and optimizes it using signal approximation (SA) [30]. Based on the complex-valued STFT spectrograms of clean speech S and noisy speech Y, the CRM is

$$\text{CRM} = \frac{Y_r S_r + Y_i S_i}{{Y_r}^2 + {Y_i}^2} + j\frac{Y_r S_i - Y_i S_r}{{Y_r}^2 + {Y_i}^2}, \tag{3}$$

where Y_r and Y_i denote the real and imaginary parts of the noisy complex spectrogram, respectively. Similarly, the real and imaginary parts of the clean complex spectrogram are represented by S_r and S_i. By multiplying the spectrogram of noisy speech $Y = Y_r + Y_i$ with the estimated mask $M = M_r + M_i$, we obtain the enhanced spectrogram in the form of: $\widetilde{S} = Y_r M_r – Y_i M_i + i(Y_r M_i + X_i M_r)$, which is converted back to the waveform using iSTFT, given by

$$\widetilde{s} = iSTFT(\widetilde{S}). \tag{4}$$

The loss function of complex spectrogram module is the well-known Scale-Invariant Signal-to-Noise Ratio (SI-SNR) [7], given below

$$\begin{cases} s_{target} := (<\widetilde{s}, s> \cdot s)/ \parallel s \parallel_2^2, \\ e_{noise} := \widetilde{s} - s_{target}, \\ \text{SI-SNR} = 10\log_{10}\left(\frac{\|s_{target}\|_2^2}{\|e_{noise}\|_2^2}\right), \end{cases} \tag{5}$$

where s and $\widetilde{s}$ are the clean and estimated time-domain waveform, respectively, $< \cdot, \cdot >$ denotes the dot product between two vectors and $\| \cdot \|_2$ is Euclidean norm (ℓ_2-norm), $\|s_{target}\|_2^2$ is the energy of the target signal, and $\|e_{noise}\|_2^2$ is the energy of the noise error. By computing the ratio of the these two and taking the logarithm, we obtain the SI-SNR value. SI-SNR measures the quality of the estimated signal by comparing the energy of the target signal to the energy of the noise error. A higher SI-SNR value indicates a higher similarity between the estimated signal and the clean signal, implying better noise removal.

C. Cascaded Architectures

With the development of both time and TF domain modules, we are now equipped to construct a cross-domain fusion model for speech enhancement. Figure 3 illustrates three distinct cascaded structures.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/TASLP.2024.3468026, IEEE/ACM Transactions on Audio Speech and Language Processing

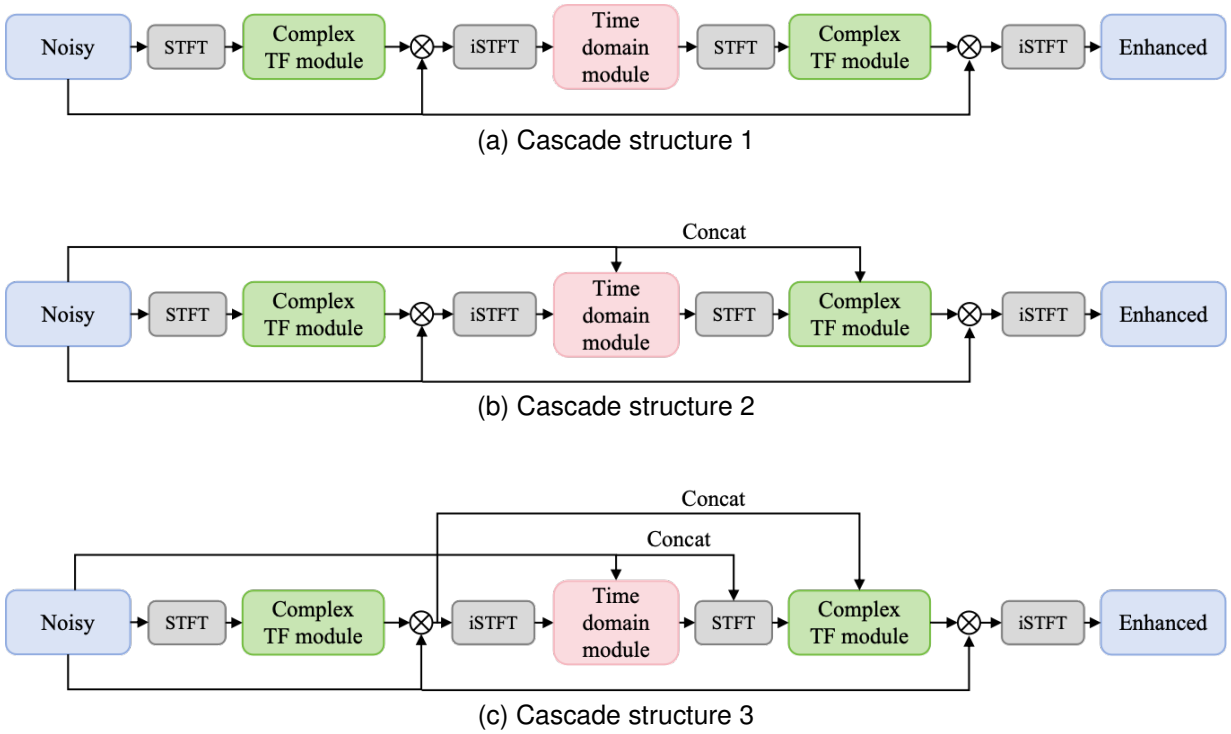


Fig. 3. Three different cascaded structures. In Cascade 1, three modules are simply concatenated, with the output of each module serving as the input to the next module. In Cascade 2, original noisy is concatenated along the channel dimension with the original input from Cascade 1. In Cascade 3, the output of the first module is concatenated along the channel dimension with the input of the third module from Cascade 2.

TABLE II
DETAILED PARAMETERS FOR COMPLEX TF MODULE, WHERE THE HYPERPARAMETERS FOR THE ENCODERS AND DECODER ARE ORDERED IN THE FASHION OF *Kernel Size, Stride, Number of Filters*, AND FOR UNIDIRECTIONAL LSTM, IT IS THE *Hidden Size*.

layer name	input size	hyperparameters	output size
$encoder_1$	$(2*N)*256*163$	$5*2,(2,1),16$	$16*128*163$
$encoder_2$	$16*128*163$	$5*2,(2,1),32$	$32*64*163$
$encoder_3$	$32*64*163$	$5*2,(2,1),64$	$64*32*163$
$encoder_4$	$64*32*163$	$5*2,(2,1),128$	$128*16*163$
$encoder_5$	$128*16*163$	$5*2,(2,1),128$	$128*8*163$
$encoder_6$	$128*8*163$	$5*2,(2,1),256$	$256*4*163$
$LSTM_1$	$256*4*163$	256	$256*4*163$
$LSTM_2$	$256*4*163$	256	$256*4*163$
$decoder_1$	$512*4*163$	$5*2,(2,1),128$	$128*8*163$
$decoder_2$	$256*8*163$	$5*2,(2,1),128$	$128*16*163$
$decoder_3$	$256*16*163$	$5*2,(2,1),64$	$64*32*163$
$decoder_4$	$128*32*163$	$5*2,(2,1),32$	$32*64*163$
$decoder_5$	$64*64*163$	$5*2,(2,1),16$	$16*128*163$
$decoder_6$	$32*128*163$	$5*2,(2,1),2$	$2*256*163$

In the first cascaded structure, we align three modules sequentially, hence its designation as a cascaded structure. This particular arrangement consists of two complex TF modules with a time-domain module sandwiched in between. The sequence of the modules is as follows: a complex TF module, followed by the time-domain module, and then another complex TF module. The first complex spectrogram module receives the noisy signal as its input. The output of this initial complex TF module then serves as the input for the ensuing time-domain module. Subsequently, the output of the time-domain module is fed into the second complex TF module. For each of these modules, specific loss functions are applied to optimize their performance. The loss function employed for the complex TF modules is SI-SNR, which is particularly suited for handling complex spectrograms. Meanwhile, the time-domain module utilizes the MSE as its loss function. This dual approach in applying loss functions is designed to effectively enhance the speech quality by addressing different aspects of signal processing inherent to each module. This cascaded structure, with its strategic sequence and tailored loss functions, aims to leverage the strengths of both time and TF domain modules for superior speech enhancement.

In cascaded structure 2, we evolve from the foundational design of cascaded structure 1 by integrating both the noisy signal and the output of the preceding module along the channel dimension. This integration serves as the input for the subsequent module. Our hypothesis is that the inclusion of the original noisy signal can significantly boost the robustness of the following modules, effectively reducing the distortion that might be introduced by the outputs of earlier modules. Consistent with cascaded structure 1, the complex spectrogram modules in this structure continue to utilize SI-SNR as their loss function, while the time-domain module retains MSE as its loss function.

Building further upon this concept, cascaded structure 3 adds an additional layer of complexity to the design of cascaded structure 2. In this structure, we incorporate the output of the first complex spectrogram module alongside the channel dimension into the input of the third module. Our rationale is

that the output from the initial module could provide valuable prior information for the third module, thereby augmenting its learning capabilities and enhancing overall model performance. The loss functions for each module in cascaded structure 3 remain consistent with the previous structures, with SI-SNR for the complex spectrogram modules and MSE for the time-domain module.

D. Parallel Architectures

The cascaded structures discussed offer a sequential approach to domain fusion, effectively combining different domains in a linear fashion. An alternative solution for domain fusion is the use of parallel structures. Figure 4 illustrates two distinct parallel structures.

In a parallel architecture, our goal is to fuse the output features of the time-domain module and the TF module in a manner that is both scientifically sound and reasonable. This fusion allows the subsequent module to access a more diverse array of speech features, which is expected to result in enhanced speech enhancement performance.

To facilitate this channel feature fusion in the parallel structure, we propose two distinct methodologies: Concatenation and Channel Attention.

1) Concatenation: This approach involves directly combining the features from both the time-domain and Complex TF modules along the channel dimension. By doing so, the model is able to concurrently process and utilize the information present in both sets of features.
2) Channel Attention [31]: In this approach, an attention mechanism is leveraged to calculate the importance or weights of each channel's features. This weighting enables the model to focus more on channels that contribute significantly to speech enhancement, thereby potentially improving the quality of the enhanced speech.

In parallel structure 1, our network architecture integrates parallel connections with a subsequent sequential connection. This structure consists of two parts: a parallel component featuring both a complex TF module and a time-domain module, and a sequential component for processing the fused output. While concatenation can merge information from different domains (Figure 4), we hypothesize that a more effective strategy would allow each domain to contribute uniquely. To achieve this, we developed an attention mechanism that assigns weights to different domains, enabling the network to autonomously determine the optimal fusion method.

The attention mechanism, depicted in Figure 5, includes several key components: Global Average Pooling (GAP) to discern the global context of channels, and a 2D convolution (Conv2d) with a kernel size of 1x1 to apprehend the local channel context. This approach uses a bottleneck structure to compute two types of channel contexts: the global channel context $\mathbf{G(X)} \in R^{C*T*F}$ and the local channel context $\mathbf{L(X)} \in R^{C*T*F}$, where C, T, and F respectively represent the dimensions of channel, time, and frequency. The calculations are

$$\mathbf{G(X)} = \beta(Conv_2(ReLU(\beta(Conv_1(\mathbf{g(X)}))))), \quad (6)$$

$$\mathbf{L(X)} = \beta(Conv_2(ReLU(\beta(Conv_1(\mathbf{X}))))), \quad (7)$$

where $Conv_1$ and $Conv_2$ represent convolution with the number of channels changing from C to $\frac{C}{r}$ and from $\frac{C}{r}$to C; β means Batch Normalization(BN); $\mathbf{g(X)}$ is the global average pooling (GAP); ReLU denotes the Rectified Linear Unit. By combining the global channel context $\mathbf{G(X)}$ and the local channel context $\mathbf{L(X)}$, the refined feature $\hat{\mathbf{X}}$ is computed as

$$\hat{\mathbf{X}} = \mathbf{X} \otimes Sig(\mathbf{G(X)} \oplus \mathbf{L(X)}), \quad (8)$$

where Sig is the Sigmoid function, $\oplus$ is the broadcasting addition and $\otimes$ denotes the element-wise multiplication. The loss function for the complex TF modules is SI-SNR, while the loss function for the time-domain module is MSE.

Parallel structure 2 builds on structure 1 by integrating the STFT-transformed noisy input signal. We combine this with outputs from both the parallel complex spectrogram and time-domain networks, feeding the composite into a sequential complex TF module. As in Parallel 1, we also use concatenation and attention operations to explore and utilize information from different domains. This approach aims to enhance speech enhancement robustness and efficacy by leveraging each domain's distinct strengths.

IV. EXPERIMENTAL RESULTS

A. Dataset

The clean speech and noise datasets from the DNS Challenge (ICASSP 2021) [32] are used for our experiments. The clean speech set includes over 500 hours of clips from 2150 speakers and the noise set includes over 180 hours of clips from 150 classes and 65,000 noise clips. We generate a training set comprising 500 hours of samples and a validation set consisting of 50 hours. To make a full use of the dataset, speech and noise signals are paired randomly, and the SNR is randomly selected between -5 dB and 20 dB. It is important to note that before mixing, all speech and noise signals are randomly truncated to 10 seconds. Additionally, we utilize the test set provided by the DNS Challenge for evaluation. To further evaluate the performance of speech enhancement models on unseen datasets, we have specifically chosen two demanding noise types, babble and factory1, from the NOISEX92 [33] dataset. Our testing protocol includes five distinct SNRs: {-6, -3, 0, 3, 6} dB. For each SNR condition, we have generated 150 pairs of noisy and clean speech samples. This experimental design allows us to assess the generalization capabilities of various speech enhancement techniques under challenging and realistic conditions.

B. Experimental Settings and Baselines

For the training of our model, we employed PyTorch framework, and the optimization of the model was conducted using the Adam optimizer. We began with an initial learning rate of 0.001, which was programmed to decay by 50% whenever there was an increase in the validation loss. The training was spread over 100 epochs, and we used a batch size of 200. All audio samples in our dataset were sampled at a rate of 16 kHz.

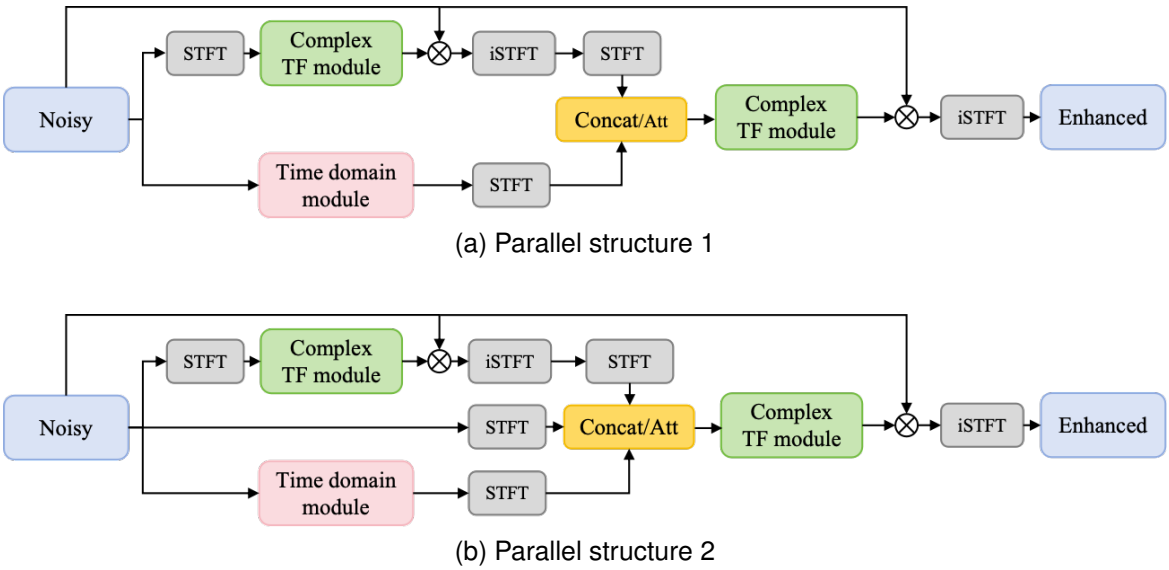


Fig. 4. Two different parallel structures. Parallel 1 consists of two components: a parallel part and a cascade part. The parallel part consists of a complex TF module and a time module. The cascade part is a complex TF module, with its input being the output of the parallel part. In parallel 2, based on parallel 1, noisy is first transformed using STFT and then concatenated along the channel dimension with the original input.

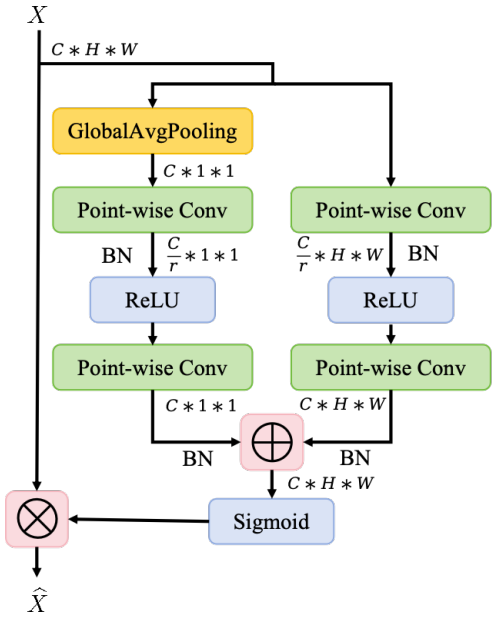


Fig. 5. The structure of Channel Attention (Att), where C represents the number of channel inputs from different modules.

To assess the performance improvement of our model, we utilized two types of Perceptual Evaluation of Speech Quality (PESQ) metrics - wide-band PESQ (WP) and narrow-band PESQ (NP) [34], along with the Short-Term Objective Intelligibility (STOI) [35] metric. WP and NP are used to evaluate the speech quality and have a range from -0.5 to 4.5, where higher values signify better quality. STOI, on the other hand, assesses speech intelligibility on a scale from 0 to 1 (100%), where higher values indicate better intelligibility, akin to the percentage of correctness.

In addition to these standard metrics, we also employed DNSMOS[1] [36] as an evaluation metric on the DNS Challenge development test set. DNSMOS is a Deep Neural Network-based non-intrusive metric that assesses speech quality.

Furthermore, we compare eight state-of-the-art speech enhancement methods using their original implementations alongside the proposed method as follows.

1) TSTNN [37]: it is a time-domain speech enhancement approach based on a two-stage Transformer network. It employs four stacked two-stage transformer blocks to progressively extract local and global information from the speech latent representation.
2) DTLN [38]: it combines a stacked-network approach that incorporates an STFT and a learned analysis and synthesis basis. This combination allows DTLN to effectively extract information from magnitude spectra while also incorporating phase information from the learned feature basis, resulting in robust speech enhancement performance.
3) TFT-Net [39]: it is a cross-domain framework that utilizes time-frequency spectra as input. It utilizes six dual-path attention blocks, which efficiently capture long-range temporal and frequency correlations while maintaining low computational costs. These blocks are responsible for generating time-domain waveforms as the output of the model in the context of speech enhancement.
4) DPT-FSNet [40]: it integrates the FullSubNet method, which combines full-band and sub-band fusion, with the DPTNet. The dual-path Transformer model in DPT-FSNet handles full-band and sub-band information through its inter and intra parts, respectively.
5) TSCN-PP [41]: it is a multi-stage model, which ranked first in 2021 DNS Challenge. Specifically, in the first

[1]https://github.com/microsoft/DNS-Challenge/tree/master/DNSMOS

stage, only the magnitude is estimated, which is then combined with the noisy phase to obtain a coarse estimation of the complex spectrum. To enhance the previous estimation, in the second stage, an auxiliary network acts as a post-processing module to further suppress residual noise and effectively modify the phase information.

6) DPRNN [42]: it employs 6 DPRNN blocks with 128 hidden units in each direction of bidirectional LSTM (BLSTM) in the time-domain audio separation network (TasNet). The TasNet consists of a linear 1-D convolutional encoder, a separator, and a linear 1-D transposed convolutional decoder.
7) FullSubNet [43]: it is a model that combines full-band and sub-band fusion to capture both full-band spectral information and long-distance cross-band dependencies. It retains the ability to model signal stationarity and attend to local spectral patterns. The full-band network consists of three LSTM layers, each with 512 hidden units, while the sub-band model includes two LSTM layers (with 384/256 units) and one dense layer.
8) DCCRN [5]: designed for complex preprocessing, this model employs complex-valued operations in both CNN and RNN structures. It uses a causal CED architecture with two complex LSTM layers between the encoder and decoder. DCCRN ranked first in the 2020 DNS Challenge.

TABLE III
SPEECH ENHANCEMENT PERFORMANCE COMPARISON ON THE DNS CHALLENGE DATASET WITHOUT REVERBERATIONS.

Model	Cau.	WP	NP	STOI	SISDR(dB)
Unprocessed	-	1.56	2.45	91.2	9.07
TSTNN	-	2.55	2.61	91.9	10.92
DPRNN	×	2.57	2.68	92.5	11.05
TFT-Net	-	2.60	2.74	92.7	11.64
DTLN	✓	-	3.04	94.7	16.34
DCCRN	✓	2.64	3.17	92.9	12.21
FullSubNet	×	2.72	3.28	95.3	16.17
DPT-FSNet	-	2.72	3.28	95.3	16.17
TSCN-PP	✓	2.94	3.42	96.6	17.99
$Cascade\ 1$	✓	2.88	3.41	96.9	19.20
$Cascade\ 2$	✓	3.08	3.54	97.4	20.20
$Cascade\ 3$	✓	3.08	3.54	97.5	20.27
$Parallel1(Cat)$	✓	3.03	3.51	97.3	19.98
$Parallel1(Att)$	✓	**3.10**	3.55	97.5	20.16
$Parallel2(Cat)$	✓	3.03	3.52	97.4	20.15
$Parallel2(Att)$	✓	**3.10**	**3.56**	**97.6**	**20.31**

C. Comparisons with Other Models

The proposed method is evaluated on the DNS challenge benchmark and compared against state-of-the-art methods. The experimental results, including the averaged SI-SDR, STOI (%), WP and NP performances, are presented in Table III. During the training stage, the noisy mixtures are generated with a random SNR ranging from -5 to 20 dB.

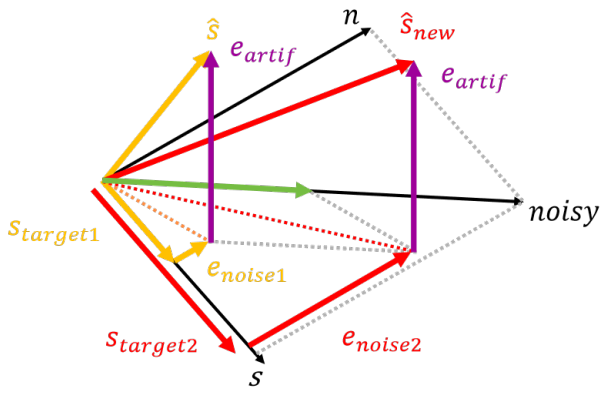


Fig. 6. Illustrations of OPD with the proposed models, where $s_{target1}$, e_{noise1}, and e_{artif} are OPD decomposition components of original signal, $s_{target2}$ and e_{noise2} are OPD decomposition components after adding the noisy input.

It can be observed from Table III that cascaded structures and parallel structures offer a better speech enhancement performance on the DNS challenge dataset. This is because our proposed cascaded and parallel models utilize sub-models from different domains for joint modeling. As a result, our models have better learning capabilities compared to single-domain models. By integrating features from different domains, our proposed models demonstrate improved generalization performance across different datasets. Specifically, the Cascade 3 and Parallel2(Att) deliver similar results, with Parallel2(Att) performing slightly better, which indicates the different roles played by different domains.

To further investigate the reasons for this improvement, we employ orthogonal projection-based error decomposition (OPD). From Figure 6, we observe that $\hat{s}$ can be decomposed into e_{noise1}, e_{artif}, and $s_{target1}$. From Table III, it is evident that Cascade 3 exhibits a significant improvement in speech quality compared to other state-of-the-art models, with WP, NP, STOI, and SI-SDR, increased by 0.14, 0.12, 0.09, and 2.28, respectively.

Analyzing Cascade 1 and Cascade 2, we can observe from Figure 3 that Cascade 2 builds upon Cascade 1 by concatenating the noisy with the inputs of the last two modules along the channel dimension. Table III shows that Cascade 2 exhibits a 0.2 improvement in WP, a 0.13 improvement in NP, a 0.5 improvement in STOI and a 1.0 improvement in SI-SDR, compared to Cascade 1. In Figure 6, we utilize the OPD to explain why introducing noisy signals leads to a significant improvement in speech quality. By combining $\hat{s}$ and noisy inputs, we obtain a new vector $\hat{s}_{new}$. From the figure, we can see that $s_{target1}/e_{noise1} > s_{target2}/e_{noise2}$, which means the ratio between s_{target} and e_{noise} decreases due to the introduction of noisy signal. However, we also have $s_{target1}/e_{artif} < s_{target2}/e_{artif}$, which indicates the share of e_{artif} in the signal decreases as well, containing less artifacts. Seeing the performance boost brought by Cascade 2, we can conclude that reducing artifacts indeed leads to a significant improvement in speech quality because e_{artif} as the unnatural signal presents the most negative impact on the final perfor-

mance. In contrast, speech quality is affected much less by adding the noise error component. This also explains why Cascade 1, despite the increasing s_{target} to e_{noise} ratio, does not outperform Cascade 2 in terms of speech enhancement performance. While Cascade 1's increased s_{target} to e_{noise} ratio comes with the cost of the introduction of e_{artif}, it is observed that artifacts have a greater impact on speech enhancement compared to noise. Consequently, the speech quality achieved after processing with Cascade 1 is lower than that of Cascade 2.

For Cascade 3, it simply concatenates the output of the first sub-model to the input of the third module, as observed in Table III, its speech enhancement performance is nearly identical to Cascade 2. This similarity is due to the fact that while incorporating the output of the first module into the input of the third module increases s_{target} to e_{noise} ratio, it also introduces the e_{artif} generated by the first module into the third module. As a result, the speech quality achieved after processing with Cascade 3 is almost the same as that of Cascade 2.

TABLE IV
DNSMOS ON THE DNS CHALLENGE DATASET WITHOUT REVERBERATIONS.

Model	Singing	Tonal	Emotion	Non-Eng	Eng	Overall
Noisy	2.96	3.00	2.67	2.96	2.80	2.86
TSCN-PP	3.14	3.44	2.92	3.50	3.49	3.38
$Cascade1$	3.15	3.51	2.99	3.56	3.65	3.42
$Cascade2$	3.15	3.54	3.00	3.59	3.67	3.42
$Cascade3$	3.15	3.55	3.00	3.60	3.69	3.43
$Parallel1(Cat)$	3.12	3.54	2.92	3.58	3.67	3.41
$Parallel1(Att)$	3.15	**3.56**	3.00	3.59	3.69	3.42
$Parallel2(Cat)$	3.13	3.55	2.93	3.58	3.68	3.41
$Parallel2(Att)$	**3.16**	3.55	**3.01**	**3.60**	**3.71**	**3.43**

We also conducted MOS evaluations for the proposed models, with MOS calculated using DNSMOS, and the test results are shown in Table IV. Furthermore, we compared the MOS results of our proposed model with the first ranked model (TSCN-PP) in 2021 DNS Challenge. From the Table III and Table IV, we can observe that the parallel models with feature fusion using Concatenation tend to slightly underperform compared to the cascaded models in terms of speech quality metrics such as PESQ, STOI, and MOS. However, the parallel structure with feature fusion using Channel Attention outperforms slightly the cascaded structure in terms of speech quality metrics. The superiority of the parallel structure with Channel Attention feature fusion method over the cascaded structure indicates that information from different domains plays distinct roles in speech enhancement models. By employing a proper fusion approach, the model can learn a wider range of speech features. Additionally, when comparing Parallel 1 to Parallel 2, the latter achieves slightly better results because it incorporates the noisy input into the input of the final model. This finding supports our earlier analysis that introducing the noisy input helps reduce the dominance of e_{artif}, resulting in improved speech enhancement effects. In comparison of concat and attention based parallel structures, it is evident that attention mechanism gives one the ability to automatically assign different weights to different domain representations, leading to a better learning capability.

To demonstrate the generalization ability of different models, we use the unseen noises from NOISEX92 to conduct inference and the results are presented in Table VI. This result also highlights the superior performance of the parallel structure with Channel Attention feature fusion in terms of speech quality metrics such as NP and STOI, in unseen noise conditions, demonstrating the model generalization capability. It is evident that the attention mechanism provides the ability to assign varying weights to different domain representations automatically, thereby enhancing the model's learning capability.

To further investigate how the time module and the complex TF module interact in speech enhancement, we display the feature outputs of the time module and the complex TF modules in both Cascade 3 and Parallel 2 structures for analysis. Figure 7 illustrates the spectrograms of the first complex TF module, time module, and second complex TF module in the order of the structure. From the representations, it is seen that the time module and complex TF module exhibit differences in learning speech features. The time module presents a richer harmonic characteristics. On the other hand, the complex TF module demonstrates a better discrimination between speech and noise. However, the complex TF module introduces significant distortions in the frequency domain, resulting in a loss of harmonic details. From the spectrogram of the second complex TF module, by fusing the output features of both the time module and the first complex TF module, the second complex TF module produces a spectrogram with enhanced frequency harmonics compared to the first complex TF module. This richer representation of speech features in the frequency domain results in a more refined output. Furthermore, the second complex TF module better discriminates between speech and noise, yielding a cleaner speech output. This showcases the different domain information indeed complement each other, producing a better harmonics representation while suppressing the noise.

Finally, we tested the enhanced audios with Whisper (small) model to evaluate the word error rate (WER) performance and results are in Table V. It is expected that the enhanced audios generally did not improve the ASR performance and even degraded the WER due to the distortions introduced. However, to further enhance ASR performance, it is advised to retrain/finetune the ASR system with the denoised data.

D. Ablation Studies

To verify the design choices of the proposed cross domain concept, in this section, we conducted a series of ablation studies to demonstrate the performance.

We first conducted an experiment to verify the performance gain of the cross-domain approach over the single-domain approach, and the results are provided in Table VII. As observed, the proposed structures, either Cascade or Parallel, deliver superior performance compared to the single-domain approach, indicating the benefits of cross-domain fusion. However, we also want to ensure that the performance gain is not

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/TASLP.2024.3468026, IEEE/ACM Transactions on Audio Speech and Language Processing

TABLE V
COMPARISON OF SPEECH RECOGNITION PERFORMANCE OF DIFFERENT MODELS.

Model	Noisy	DPRNN	DCCRN	FullSubNet	Cascade 3	Parallel2(Att)
WER	23.5	31.2	27.7	28.1	25.3	25.1

TABLE VI
OBJECTIVE RESULT COMPARISONS AMONG DIFFERENT MODELS IN TERMS OF NP AND STOI FOR THE UNSEEN BABBLE AND FACTORY1 NOISES.

	Model	NP						STOI					
	SNR	-6	-3	0	3	6	Avg.	-6	-3	0	3	6	Avg.
Babble	Noisy	1.49	1.64	1.82	2.01	2.23	1.84	24.33	31.51	39.66	48.21	57.74	40.29
	LSTM	1.72	1.98	2.25	2.46	2.67	2.22	41.97	52.33	61.63	68.89	75.60	60.09
	CRN [23]	1.67	2.01	2.31	2.56	2.80	2.27	42.63	53.82	63.48	71.20	78.01	61.83
	GCRN [44]	1.88	2.25	2.58	2.83	3.05	2.52	48.41	59.77	69.09	75.78	80.87	66.78
	DCCRN	1.84	2.21	2.55	2.81	3.07	2.49	45.86	57.71	67.61	75.39	81.51	65.62
	ConvTasNet [7]	1.89	2.21	2.50	2.73	2.94	2.45	53.87	64.37	72.61	78.63	83.49	70.59
	TSCN-PP	2.10	2.51	2.83	3.06	3.26	2.75	56.75	68.57	76.35	81.89	85.88	73.89
	$Parallel2(Cat)$	2.17	2.55	2.91	3.12	3.29	2.80	57.91	70.67	79.32	83.67	87.52	75.81
	$Parallel2(Att)$	**2.33**	**2.67**	**2.95**	**3.18**	**3.37**	**2.90**	**60.13**	**73.26**	**81.25**	**85.37**	**90.65**	**78.13**
Factory1	Noisy	1.36	1.55	1.75	1.96	2.17	1.76	23.67	31.97	41.13	50.32	59.78	41.37
	LSTM	1.83	2.12	2.36	2.56	2.73	2.32	42.13	53.50	62.75	70.09	75.89	60.87
	CRN	1.84	2.16	2.42	2.66	2.86	2.39	42.39	54.21	64.03	71.92	78.18	62.15
	GCRN	2.00	2.39	2.68	2.90	3.09	2.61	45.73	59.37	68.85	75.70	80.52	66.03
	DCCRN	2.07	2.42	2.70	2.93	3.13	2.65	46.73	59.50	68.92	76.16	81.81	66.62
	ConvTasNet	2.02	2.32	2.56	2.79	2.99	2.54	51.72	63.48	72.07	78.36	82.97	69.72
	TSCN-PP	2.29	2.62	2.86	3.09	3.25	2.82	56.50	67.90	75.49	81.25	85.12	73.25
	$Parallel2(Cat)$	2.30	2.67	2.91	3.12	3.29	2.85	59.01	68.16	77.23	83.21	87.27	74.97
	$Parallel2(Att)$	**2.31**	**2.76**	**2.97**	**3.15**	**3.37**	**2.91**	**65.05**	**71.19**	**80.92**	**85.31**	**92.79**	**79.05**

TABLE VII
PERFORMANCE COMPARISON OF CROSS DOMAIN APPROACHES OVER SINGLE DOMAIN, WHERE MACS ARE CALCULATED BY FUNCTION 'PROFILE'.

Model	WP	NP	STOI	SISDR(dB)	Para. (M)	MACs(G/s)
$Time\ module$	2.21	2.65	92.1	11.10	10.13	2.45
$TF\ module$	2.66	3.21	93.2	12.27	2.66	3.21
$TF\ module+$	2.81	3.36	96.5	18.34	22.42	15.05
$Cascade\ 1$	2.88	3.41	96.9	19.20	17.94	11.29
$Cascade\ 2$	3.08	3.54	97.4	20.20	17.94	11.33
$Cascade\ 3$	3.08	3.54	97.5	20.27	17.94	11.36
$Parallel1(Cat)$	3.03	3.51	97.3	19.98	17.94	11.32
$Parallel1(Att)$	**3.10**	3.55	97.5	20.16	18.19	11.75
$Parallel2(Cat)$	3.03	3.52	97.4	20.15	17.94	11.36
$Parallel2(Att)$	**3.10**	**3.56**	**97.6**	**20.31**	18.19	11.79

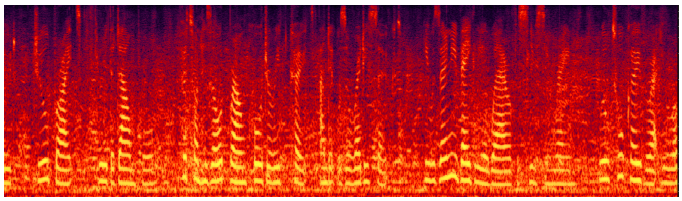
TABLE VIII
PERFORMANCE COMPARISON OF DIFFERENT POSITIONS OF THE TIME DOMAIN MODULE.

Model	Position	WP	NP	STOI	SISDR(dB)
$Cascade\ 3$	Beginning	3.03	3.55	97.2	20.20
$Cascade\ 3$	Middle	3.08	3.54	97.5	20.27

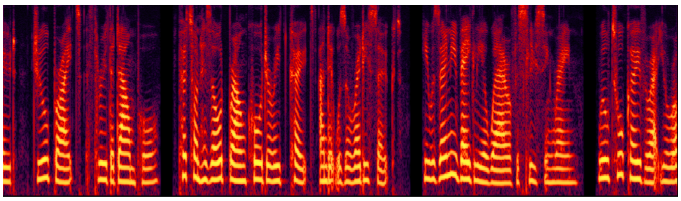
merely due to the increased model parameters. To this end, we increased the model size of the TF domain module, termed as *TF module+*, to match that of the cross-domain models. The results show that while increasing the model size does improve the metrics, its performance still falls short of the proposed models. This shortfall is attributed to the lack of complementary information provided by different domains.

To investigate the influence of the order of sub-modules, in Cascade 3, we place the time module at the beginning instead of in the middle, as in our current design. The results, provided in Table VIII, show that both designs yield similar performance. This indicates that cross-domain optimization can effectively fuse information from different domains regardless of the time module's position. However, placing the time module in the middle does offer a slight performance improvement. In addition, we are also able to provide evaluations of different

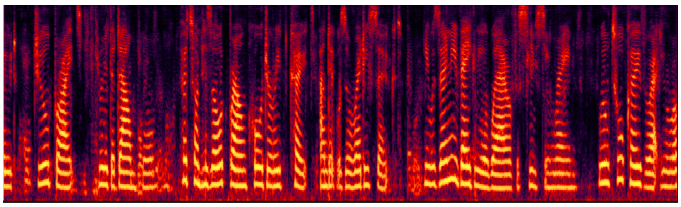
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/TASLP.2024.3468026, IEEE/ACM Transactions on Audio Speech and Language Processing



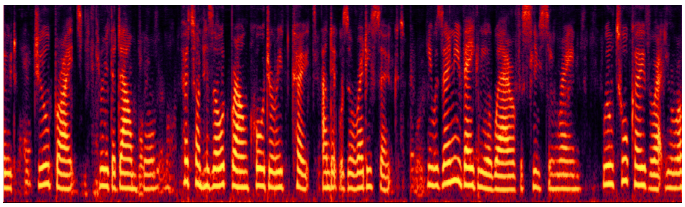
(a) Spectrogram of noisy signal (WP=1.56, NP=1.45, STOI=91.2, SDR=9.07)



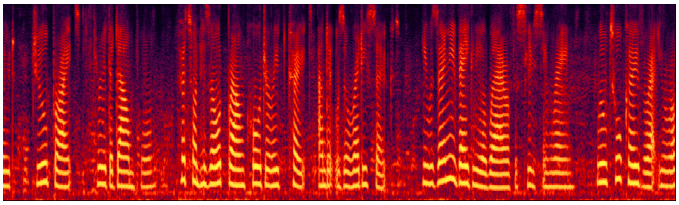
(b) Spectrogram of clean signal



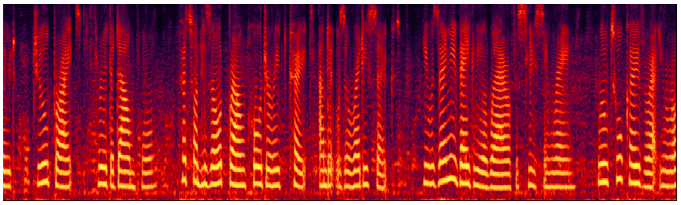
(c) Spectrogram of Cascade 3 (WP=2.82, NP=3.37, STOI=96.6, SDR=18.34)



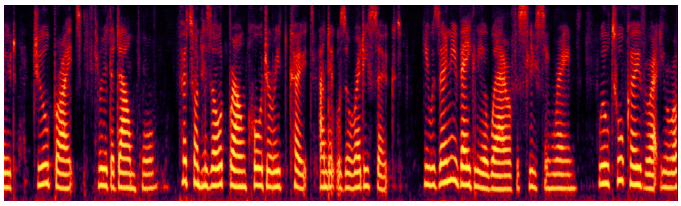
(d) Spectrogram of Parallel 2 (WP=2.78, NP=3.35, STOI=92.0, SDR=18.35)



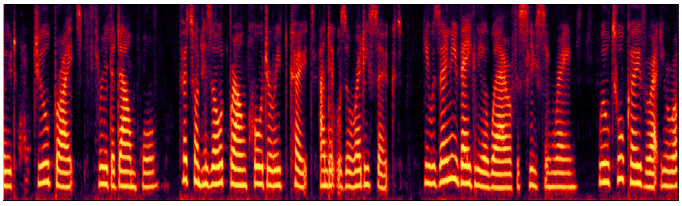
(e) Spectrogram of Cascade 3 (WP=2.25, NP=3.01, STOI=94.2, SDR=12.65)



(f) Spectrogram of Parallel 2 (WP=2.30, NP=3.15, STOI=94.7, SDR=12.71)



(g) Spectrogram of Cascade 3 (WP=3.08, NP=3.54, STOI=97.5, SDR=20.27)



(h) Spectrogram of Parallel 2 (WP=3.10, NP=3.56, STOI=97.6, SDR=20.31)

Fig. 7. The different layer representations of different modules from different structures. First row: Noisy and clean speeches; Second row: Output of first complex TF module; Third row: Output of time domain module; Fourth row: Output of second complex TF module.

stages to see the impact of each sub-module, summarized in Table IX. One interesting phenomenon is that the performance of the time module is usually worse than that of the TF module, likely due to the inefficient learning capacity of the time module. This observation is consistent with the output in Figure 7. However, the final performance is improved by effectively fusing information from different domains. Overall, it is concluded that the final performance is less dependent on the order of each module, and each stage may not necessarily generate incremental gains over the previous stage. Instead, the final performance improvement is primarily due to the fusion of cross-domain information, highlighting the significance of our design in leveraging multiple-domain information.

In our designs, we also leverage the original noisy inputs to be used in the system. To show the benefits of doing so, we conducted another ablation study where we add different levels of original noisy inputs in Cascade structure, i.e., $\hat{s} + w * y$ with $w = 0, 0.5, 1$, where y is the noisy speech. Note that when $w = 0$, it is Cascade 1, and when $w = 1$, it is Cascade 3. From Table X, it is seen that gradually adding original

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/TASLP.2024.3468026, IEEE/ACM Transactions on Audio Speech and Language Processing

TABLE IX
PERFORMANCE METRICS OF DIFFERENT STAGES.

Model	TF module 1				Time module				TF module 2			
Metrics	WP	NP	STOI	SDR	WP	NP	STOI	SDR	WP	NP	STOI	SDR
Cascade 3	2.82	3.37	96.6	18.34	2.25	3.01	94.2	12.65	3.08	3.54	97.5	20.27
Parallel 2(Att)	2.78	3.35	92.0	18.35	2.30	3.15	94.7	12.71	3.10	3.56	97.6	20.31

TABLE X
PERFORMANCE COMPARISON OF ADDING DIFFERENT LEVELS OF ORIGINAL NOISY INPUTS.

w	WP	NP	STOI	SISDR(dB)
0	2.88	3.41	96.9	19.20
0.5	2.91	3.47	97.1	19.93
1	3.08	3.54	97.5	20.27

noisy input indeed improves the final performance due to the increased combination of clean and noisy speech features and the decreased noise artifact e_{artif}. As we increase the weight w, the performance metrics WP, NP, STOI, and SISDR all show a noticeable improvement. Specifically, at $w = 1$, the WP improves to 3.08, NP to 3.54, STOI to 97.5, and SISDR to 20.27 dB, indicating that incorporating the noisy input effectively enhances the overall performance. This result is consistent with the our early OPD analysis.

V. Conclusion

In this work, to utilize different domain information, we developed five structures to investigate how each domain contributes to the final results. First, we design standalone time domain and complex TF domain modules for fusion purposes. Second, to leverage cross-domain information, we mainly design cascade and parallel structures, three cascade and two parallel structures, to be exact. The experimental results show that the developed Cascade 3 and Parallel 2 with attention produces superior results, indicating that information flows provide more complementary ability to each other and attention is able to determine which domain learns a better representation. Compared with other methods, the proposed models outperform them, suggesting that leveraging multiple domains indeed benefits the final performance. In our current study, the cascade and parallel structures are manually designed, however, in the future, we would like to explore the possibility of using neural architecture search (NAS) to identify the best model given the dataset and constraints.

References

[1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.

[2] Y. Sun, W. Wang, J. Chambers, and S. M. Naqvi, "Two-stage monaural source separation in reverberant room environments using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 125–139, 2018.

[3] S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "A recurrent variational autoencoder for speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 371–375.

[4] Y. Sun, Y. Xian, W. Wang, and S. M. Naqvi, "Monaural source separation in complex domain with long short-term memory neural network," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 359–369, 2019.

[5] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *INTERSPEECH*, 2020, pp. 2427–2476.

[6] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[7] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[8] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1570–1584, 2018.

[9] D. Stoller, S. Ewert, and S. Dixon, "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, pp. 334–340.

[10] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[11] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 708–712.

[12] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," *Speech Separation by Humans and Machines*, pp. 181–197, 2005.

[13] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[14] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2015.

[15] K. Tan and D. Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6865–6869, 2019.

[16] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *International Conference on Machine Learning (ICML)*, 2009, pp. 41–48.

[17] E. Tzinis, S. Venkataramani, Z. Wang, C. Subakan, and P. Smaragdis, "Two-step sound source separation: Training on learned latent targets," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 31–35.

[18] J. Lin, A. J. d. L. van Wijngaarden, K.-C. Wang, and M. C. Smith, "Speech enhancement using multi-stage self-attentive temporal convolutional networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3440–3450, 2021.

[19] L. Zhang, M. Wang, A. Li, Z. Zhang, and X. Zhuang, "Incorporating multi-target in multi-stage speech enhancement model for better generalization," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2021, pp. 553–558.

[20] H. Wang and D. Wang, "Cross-domain speech enhancement with a neural cascade architecture," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7862–7866.

[21] J. Andreas, J. H. Eric, M. Nicola, B. Rachel, K. Aparna, and W. Tillman, "Singing voice separation with deep U-Net convolutional networks," *In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 323–332, 2017.
[22] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241.
[23] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *INTERSPEECH*, 2018, pp. 3229–3233.
[24] 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," in *INTERSPEECH*, 2020, pp. 11 531–11 539.
[25] X. Hao, X. Su, S. Wen, Z. Wang, Y. Pan, F. Bao, and W. Chen, "Masking and inpainting: A two-stage speech enhancement approach for low snr and non-stationary noise," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6959–6963.
[26] A. Li, W. Liu, C. Zheng, C. Fan, and X. Li, "Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1829–1843, 2021.
[27] K. Iwamoto, T. Ochiai, M. Delcroix, R. Ikeshita, H. Sato, S. Araki, and S. Katagiri, "How bad are artifacts?: Analyzing the impact of speech enhancement errors on ASR," in *INTERSPEECH*, 2022, pp. 5418–5422.
[28] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.
[29] R. Gu, J. Wu, S.-X. Zhang, L. Chen, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, *End-to-end multi-channel speech separation*. arXiv preprint arXiv:1905.06286, 2019.
[30] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1026–1034.
[31] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 531–11 539.
[32] C. K. Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "INTERSPEECH 2021 Deep Noise Suppression Challenge," in *INTERSPEECH*, 2021, pp. 2796–2800.
[33] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.
[34] A. W. Rix, J. G. Beerends, M. P. Hollier, and 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)*, 2001, pp. 749–752.
[35] D. Wu, B. Zhang, C. Yang, Z. Peng, W. Xia, X. Chen, and X. Lei, "U2++: Unified two-pass bidirectional end-to-end model for speech recognition," *arXiv preprint arXiv:2106.05642*, 2021.
[36] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6493–6497.
[37] K. Wang, B. He, and W.-P. Zhu, "TSTNN: Two-stage transformer based neural network for speech enhancement in the time domain," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7098–7102.
[38] N. L. Westhausen and B. T. Meyer, "Dual-signal transformation LSTM network for real-time noise suppression," in *INTERSEECH*, 2020, pp. 2477–2481.
[39] C. Tang, C. Luo, Z. Zhao, W. Xie, and W. Zeng, "Joint time-frequency and time domain learning for speech enhancement," in *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 2021, pp. 3816–3822.
[40] F. Dang, H. Chen, and P. Zhang, "DPT-FSNet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6857–6861.
[41] A. Li, W. Liu, X. Luo, C. Zheng, and X. Li, "ICASSP 2021 deep noise suppression challenge: Decoupling magnitude and phase optimization with a two-stage deep network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6628–6632.
[42] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-Path RNN: efficient long sequence modeling for time-domain single-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 46–50.
[43] X. Hao, X. Su, R. Horaud, and X. Li, "FullSubNet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6633–6637.
[44] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 380–390, 2019.