Neural Cascade Architecture With Triple-Domain Loss for Speech Enhancement

Heming Wang, *Student Member, IEEE*, and DeLiang Wang, *Fellow, IEEE*

***Abstract*—This paper proposes a neural cascade architecture to address the monaural speech enhancement problem. The cascade architecture is composed of three modules which optimize in turn enhanced speech with respect to the magnitude spectrogram, the time-domain signal and the complex spectrogram. Each module takes as input the noisy speech and the output obtained from the previous module, and generates a prediction of the respective target. Our model is trained in an end-to-end manner, using a triple-domain loss function that accounts for three domains of signal representation. Experimental results on the WSJ0 SI-84 corpus show that the proposed model outperforms other strong speech enhancement baselines in terms of objective speech quality and intelligibility.**

***Index Terms*—Monaural speech enhancement, time domain, complex domain, cascade architecture, deep learning.**

I. Introduction

SPEECH enhancement attempts to remove background noise from the speech signal in a noisy environment, in order to improve the intelligibility and quality of the noisy speech. It is extensively applied in speech processing tasks, such as automatic speech recognition, telecommunication, and hearing prosthesis. In this paper we study monaural speech enhancement, where noisy speech is collected from a single microphone.

Traditional approaches to monaural speech enhancement include spectral subtraction and statistical estimation [19], as well as computational auditory scene analysis [41]. Recently, supervised algorithms based on deep neural networks (DNNs) have been established as the mainstream approach [42]. Popular networks include recurrent neural networks (RNNs) [2], [20], [34], [45], convolutional neural networks (CNNs) [7], [23], [28], and generative adversarial networks [4], [29], [31]. Early DNN studies focus on the magnitude spectrogram of noisy speech in the time-frequency (T-F) domain, which is derived from short-time Fourier transform (STFT), and leaves the phase of noisy speech unaltered. The training targets of these studies can be categorized into two groups. One group consists of masking based targets such as the ideal binary mask [40] and the ideal ratio mask (IRM) [43], and the other group includes mapping based targets like the target magnitude spectrum [9], [47]. Recent works in speech enhancement emphasize the importance of phase estimation motivated by the observation that accurate phase estimation leads to a significant improvement in speech quality [22]. To this end, complex-domain and time-domain approaches have been proposed to address both magnitude and phase estimation. Based on the insight that real and imaginary spectrograms both exhibit T-F structure whereas phase spectrogram does not [46], complex T-F masking and spectral mapping aim to recover the phase information by estimating real and imaginary components simultaneously [3], [6], [12], [37], [46]. Other studies tackle this problem in the time domain by directly estimating waveform signals [21], [26], [27], [29], such that the phase can be implicitly estimated. Other studies have attempted to reconstruct clean speech using cross-domain techniques. For instance, Pandey and Wang [26] train a time-domain autoencoder but optimize the prediction with respect to the magnitude spectrogram by applying STFT to the predicted waveform, which substantially improves objective speech quality. Bahmaninezhad *et al.* [1] incorporate a time-domain scale-invariant source-to-noise ratio in the separation criterion of frequency-domain speech separation, and their experiments show the advantage of the cross-domain model over the same-domain counterpart.

Despite the success of recent studies that take phase information into consideration, jointly enhance magnitude and phase in one stage could be difficult, especially under very low signal-to-noise ratio (SNR) conditions for highly non-stationary noises. In contrast to these single-stage models, multi-stage networks decompose a difficult task into easier sub-tasks. This strategy has been extensively applied in the speech field. For example, Zhao *et al.* [50] perform noisy and reverberant speech enhancement in two stages, where the first stage deals with additive noise and the second stage convolutive reverberation. In the two-stage network of Hao *et al.* [10], the first stage predicts binary masks in the T-F spectrogram to remove T-F units that are dominated by noise, and then a CNN is trained to perform inpainting in order to recover the masked magnitude spectrogram. Li *et al.* [16] use a two-stage network to progressively recovers the clean speech from a noisy mixture. During the first-stage training, the first subnetwork produces a coarse magnitude estimate. In the next stage, the second subnetwork conducts complex spectral mapping and is jointly trained with the first-stage module. Compared with directly estimating the complex spectrogram, using estimated magnitude spectrogram

Manuscript received July 25, 2021; revised November 15, 2021; accepted December 7, 2021. Date of publication December 28, 2021; date of current version February 9, 2022. This work was supported in part by NIDCD under Grant R01 DC012048 and in part by Ohio Supercomputer Center. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiao-Lei Zhang. *(Corresponding author: DeLiang Wang.)*

Heming Wang is with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: wang.11401@osu.edu).

DeLiang Wang is with the Department of Computer Science and Engineering, and the Center for Cognitive and Brain Sciences, The Ohio State University, Columbus 43210 USA (e-mail: dwang@cse.ohio-state.edu).

Digital Object Identifier 10.1109/TASLP.2021.3138716

3

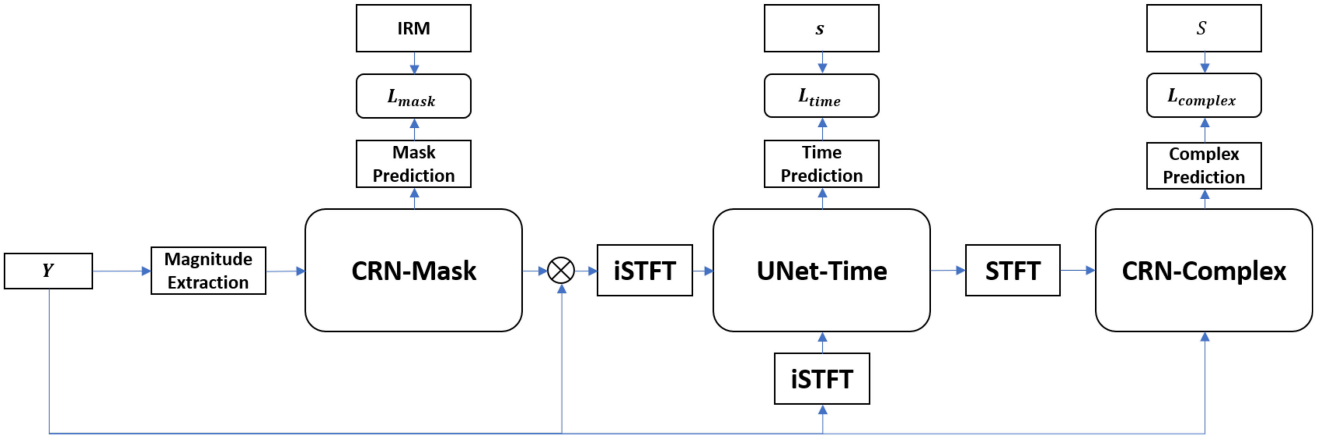


Fig. 1. A schematic diagram illustrating the cascade architecture. The first module employs a CRN to predict T-F masks, followed by a UNet to predict time-domain signals. The last module also employs CRN to operate on the complex spectrogram, and its output represents the outcome of the proposed network.

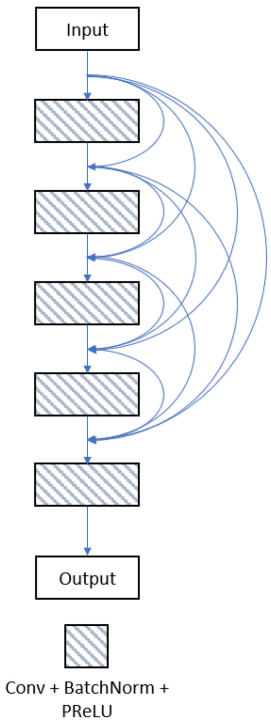


Fig. 2. Densely Connection Block.

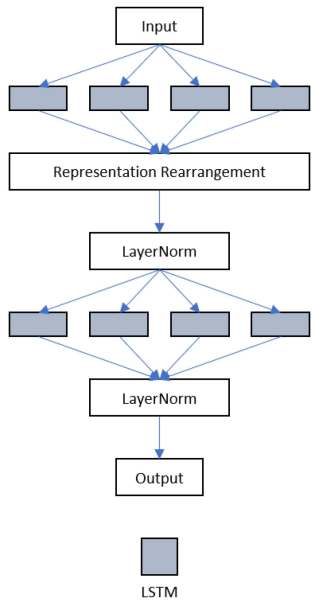


Fig. 3. Grouped Two-layer LSTMs.

as additional input considerably improves the enhancement performance. Tzinis *et al.* [38] present a two-step training procedure for sound source separation. The first step learns a latent space representation for the input audio, and the second step utilizes a separation module to separate in the learned latent space. They show that a pretrained optimal latent space results in consistent improvement for sound separation tasks. In [5], Fan *et al.* propose a multi-stage network for speech separation. It contains three stages. A pre-separation stage utilizes a T-F domain separation method. In the next stage, a fully convolutional network uses waveform as input to further enhance the pre-separated speech. A deep attention module is incorporated for feature fusion. Finally, a temporal convolutional module (TCM) is applied to model long-term temporal dependencies. Lin *et al.* [18] apply multi-stage learning to speech enhancement by stacking TCM blocks. Each block consists of a self-attentive TCM, and the proposed network performs sequential refinement for the magnitude spectrogram. Experiments show that progressively estimating magnitude spectrogram leads to noticeable improvement, and the enhancement performance does not further improve after 5 stages of TCM blocks.

In this study, we propose a novel neural cascade architecture for monaural speech enhancement. The cascade architecture consists of three modules based on convolutional recurrent network (CRN) [36] and UNet [33]. The rationale behind the cascade design is to constrain enhanced speech in multiple domains of signal representation using a triple-domain loss.

Each enhancement module operates on the output of the previous module and the original noisy speech, such that the speech is enhanced progressively and, at the same time, allowing for correction of estimation errors of the previous module. Different from multi-stage enhancement studies that undergo multiple sequential training processes, our cascade architecture is trained only once.

The remainder of the paper is organized as follows. In Section II, we formulate the monaural speech enhancement problem. In Section III, we present the proposed cascade architecture in detail. In Section IV, we provide experimental setting and data setup. In Section V, experimental results on WSJ0 SI-84 are displayed to demonstrate the performance of our model, along with the comparisons of state of the art baselines. Finally in Section IV we conclude the paper.

II. Problem Formulation

For monaural speech enhancement, we are given a single-microphone noisy mixture y that is composed of clean target speech s and background noise n, expressed as

$$y[k] = s[k] + n[k], \tag{1}$$

where k indicates a time sample. Converting to the T-F domain by applying STFT, we have,

$$Y(t,f) = S(t,f) + N(t,f), \tag{2}$$

where Y, S and N are the corresponding STFTs of y, s and n, and t, f index time frame and frequency bin, respectively. These STFTs can be expressed in either Cartesian or polar coordinates. In Cartesian coordinates, they are expressed as the addition of real and imaginary parts,

$$\begin{aligned} Y_r(t,f) + iY_i(t,f) = (S_r(t,f) + N_r(t,f)) \\ + i(S_i(t,f) + N_i(t,f)). \end{aligned} \tag{3}$$

Here subscripts r and i indicate real and imaginary parts of STFT, respectively, and i is the imaginary unit. From the polar coordinate perspective, STFT can be expressed as the product of magnitude and phase,

$$|Y(t,f)|e^{i\theta_Y(t,f)} = |S(t,f)|e^{i\theta_S(t,f)} + |N(t,f)|e^{i\theta_N(t,f)}, \tag{4}$$

where $|\cdot|$ denotes the magnitude and θ the phase in radius.

The goal of speech enhancement is to produce an estimate $\hat{s}$ that is close to the original clean speech s with a DNN f. As mentioned in Section I, our network has three modules f_{mask}, f_{time} and $f_{complex}$ that enhance the noisy speech from different perspectives. With the parameters of each module denoted as ϕ, we formulate the speech enhancement problem as,

$$\begin{aligned} \hat{S}_1(t,f) &= f_{mask}(\phi_{mask}, |Y|(t,f)) \odot Y(t,f) \\ \hat{s}_2(k) &= f_{time}(\phi_{time}, y(k), \hat{s}_1(k)) \\ \hat{S}_3(t,f) &= f_{complex}(\phi_{complex}, Y(t,f), \hat{S}_2(t,f)), \end{aligned} \tag{5}$$

where the subscript number 1,2,3 indicates the module number, and $\odot$ element-wise multiplication. The final enhancement result is the output of the last module, i.e. $\hat{S}_3$.

III. Cascade Architecture

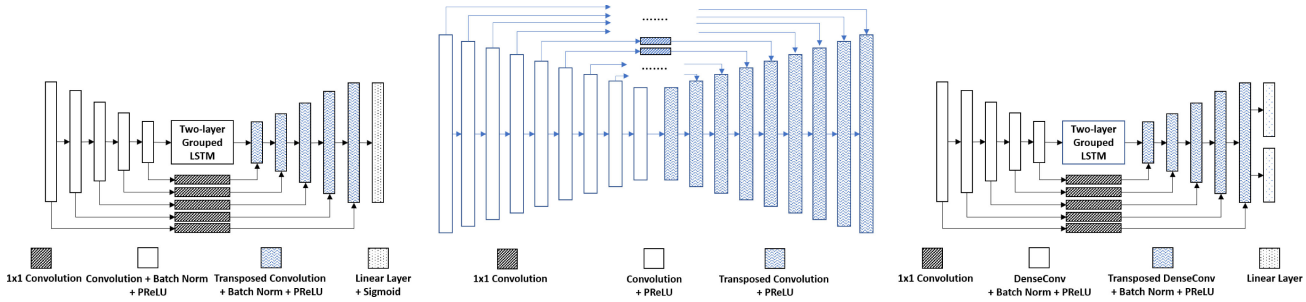
The proposed network for speech enhancement is shown in Fig. 1. The cascade architecture consists of three modules: the magnitude mask module CRN-Mask, the time-domain module UNet-Time, and the complex-domain module CRN-Complex. Each module operates on the noisy speech input, and the result of the previous module. In addition, each module generates an output that will be optimized directly during DNN training. The input to the cascade architecture is the complex spectrogram of a noisy mixture. CRN-Mask is fed with the noisy magnitude spectrogram, and estimates the IRM. The estimated mask is then multiplied with the original complex spectrogram to provide the magnitude-masked input to the next module. By applying inverse STFT (iSTFT), the noisy complex input and magnitude-masked input are converted to waveform signals and are fed to UNet-Time. The second module, UNet-Time, produces a time-domain estimate of the clean speech, which is then converted back to the T-F domain via STFT. The last module, CRN-Complex, takes the original input and the output of UNet-Time to perform complex spectral mapping. In the following subsections, we first introduce key components employed in our model, then describe network configurations and training objectives for the modules.

A. Densely-Connected Convolutional Block

Inspired by the recently proposed densely connected convolutional network [13], [25], we introduce dense connections to our design. As shown in Fig. 2, we utilize a densely-connected (DC) convolutional block to replace a standard convolution in the complex-domain module. The DC block is designed based on the idea of reusing feature maps by decomposing one convolution layer into several with fewer channels, and densely connecting these layers. Such a pattern improves the information flow between layers as they are all directly connected. Specifically, our dense block has 5 layers, each of which consists of a 2D convolution, a batch normalization layer and a parametric rectified linear unit (PReLU) activation function [11]. The growth rate for the dense block is set to 8, meaning that the number of the output channels of the first four convolutional layers is 8. The final layer accepts all previous outputs and performs the normal convolution operation.

B. Grouping Strategy for RNN

To reduce the computational complexity and the number of trainable parameters of the cascade model, we employ the grouped long short-term memory (LSTM) proposed by Gao *et al.* [8], which applies a grouping strategy to improve the efficiencies of RNN computations. Specifically, for recurrent layers, we split the features into disjoint groups to reduce the number of inter-layer connections. We also rearrange the representations between two successive recurrent layers to model the intra-group dependency, as suggested in [37]. This is shown in Fig. 3, and we utilize a group of 4 and apply a layer normalization after each LSTM layer. In practice we find that this technique improves



(a) CRN-Mask (b) UNet-Time (c) CRN-Complex

Fig. 4. Illustration of the modules of the neural cascade architecture. From left to right are the mask module CRN-Mask, the time module UNet-Time, and the complex module CRN-Complex.

the computational efficiency while maintaining enhancement performance.

C. Mask Module and Complex Module

We use a standard CRN to construct the mask module [36], which is fed with a noisy magnitude spectrogram and predicts the IRM; we have tried directly estimating magnitudes and it does not perform as well as IRM estimation. As illustrated in Fig. 4(a), CRN is an encoder-decoder structure network that uses LSTMs in the bottleneck to model temporal dependencies. We utilize a stack of 5 convolutional layers with a stride 2 for both encoder and decoder, except that the encoder uses normal convolutions to downsample along the frequency axis and the decoder uses transposed convolutions for upsampling. As such, the encoder-decoder structure is symmetric. For skip connections we use the pointwise convolution to concatenate the output of each encoder layer to the corresponding decoder layer, which our experiments show to outperform the simple concatenation. Each convolution is followed by batch normalization, and a PReLU non-linearity. The final mask prediction is generated by appending a linear layer with a sigmoidal activation function. The complex module is depicted in Fig. 4(c) and it resembles the mask module. The major differences are two-fold. First, all convolutional layers in the encoder and decoder are replaced with densely connected blocks. Second, the output of the decoder is split into two halves and each half is reshaped into 1D and followed by a linear layer to generate real and imaginary estimates separately. No non-linearities are used for the final layers.

D. Time Module

The time module UNet-Time is an encoder-decoder structure that is based on the standard UNet. It operates in the time domain and enhances frame-level speech segments. As illustrated in Fig. 4(b), the time module is a fully convolutional network comprising 9 convolutional layers for both the encoder and the decoder. Similarly, we use pointwise convolutions for skip connections, and PReLU to provide non-linearity. We do not perform batch normalization as it does not provide performance benefits in our experiments. Note that we do not use dense blocks in the time module as the complexity will drastically increase.

Table I summarizes the network design details of individual modules. Since CRN-Mask and CRN-Complex are similar, we only present the parameter setup of the mask module for brevity. Layer names denote the function and position of the corresponding layer or block. The input and output size of each layer are marked as $Channels \times TimeSteps \times FreqChannels$. Additionally, the hyperparameters for each layer are specified in the format of $KernelSizes$, $Strides$, $OutChannels$.

E. Training Targets and Loss Functions

The training objective of our cascade architecture is composed of three parts, corresponding to the outputs generated by the three modules. As mentioned in Section I, the widely used training target of the IRM [43] is defined in the magnitude domain. Specifically, it is based on the energy of speech and noise in T-F units,

$$\mathrm{IRM}(t,f) = \sqrt{\frac{S(t,f)^2}{S(t,f)^2 + N(t,f)^2}}. \tag{6}$$

The first loss L_{mask} is calculated based on IRM estimation,

$$L_{mask} = \frac{1}{TF}\sum_{t,f} |\mathrm{RM}(t,f) - \mathrm{IRM}(t,f)|, \tag{7}$$

where RM denotes the predicted ratio mask, and T and F represent the number of time frames and frequency bins, respectively. This loss corresponds to the mean absolute error between the predicted ratio mask and the IRM.

The second loss L_{time} is motivated by the phase-constrained magnitude loss introduced in [24], which takes into consideration the STFT magnitudes of both speech and noise. Specifically,

$$\begin{aligned} L_{time} = {} & \frac{1}{TF}\sum_{t,f}(|\hat{S}(t,f)| - |S(t,f)|) \\ & + \frac{1}{TF}\sum_{t,f}(|Y(t,f) - \hat{S}(t,f)| - |N(t,f)|), \end{aligned} \tag{8}$$

where $\hat{S}$ is the STFT of the predicted waveform speech, and $Y - \hat{S}$ is the STFT of the estimated noise. This loss is demonstrated to be effective for imposing a phase constraint on optimization, and leads to good objective quality scores.

TABLE I
NETWORK DETAILS OF CRN-MASK AND UNET-TIME

CRN-Mask				UNet-Time			
layer name	input size	hyperparameters	output size	layer name	input size	hyperparameters	output size
conv2d_1	$1 \times T \times 161$	(1,4), (1,2), 12	$12 \times T \times 80$	conv1d_1	1×2048	11, 2, 20	20×1024
conv2d_2	$12 \times T \times 80$	(1,4), (1,2), 24	$24 \times T \times 40$	conv1d_2	20×1024	11, 2, 40	40×512
conv2d_3	$24 \times T \times 40$	(1,4), (1,2), 48	$48 \times T \times 20$	conv1d_3	40×512	11, 2, 60	60×256
conv2d_4	$48 \times T \times 20$	(1,4), (1,2), 96	$96 \times T \times 10$	conv1d_4	60×256	11, 2, 80	80×128
conv2d_5	$96 \times T \times 10$	(1,4), (1,2), 192	$192 \times T \times 5$	conv1d_5	80×128	11, 2, 100	100×64
reshape_1	$192 \times T \times 5$	-	$12 \times T \times 80$	conv1d_6	100×64	11, 2, 120	120×32
grouped_lstm_1	$T \times 960$	960	$T \times 960$	conv1d_7	120×32	11, 2, 140	140×16
grouped_lstm_2	$T \times 960$	960	$T \times 960$	conv1d_8	140×16	11, 2, 160	160×8
reshape_2	$192 \times T \times 5$	-	$12 \times T \times 80$	conv1d_9	160×8	11, 2, 180	180×4
transconv2d_5	$384 \times T \times 5$	(1,3), (1,2), 96	$192 \times T \times 10$	transconv1d_9	180×4	11, 2, 180	180×8
transconv2d_4	$192 \times T \times 10$	(1,3), (1,2), 48	$96 \times T \times 20$	transconv1d_8	340×8	11, 2, 160	160×16
transconv2d_3	$96 \times T \times 20$	(1,3), (1,2), 24	$48 \times T \times 40$	transconv1d_7	300×16	11, 2, 140	140×32
transconv2d_2	$48 \times T \times 40$	(1,3), (1,2), 12	$24 \times T \times 80$	transconv1d_6	260×32	11, 2, 120	120×64
transconv2d_1	$24 \times T \times 80$	(1,3), (1,2), 1	$1 \times T \times 161$	transconv1d_5	220×64	11, 2, 100	100×128
linear	$1 \times T \times 161$	161	$1 \times T \times 161$	transconv1d_4	180×128	11, 2, 80	80×256
				transconv1d_3	140×256	11, 2, 60	60×512
				transconv1d_2	100×512	11, 2, 40	40×1024
				transconv1d_1	60×1024	11, 2, 20	20×2048
				out_conv	20×2048	1, 1, 1	1×2048

The last loss $L_{complex}$ is defined in terms of the complex spectrogram, and it combines the real and imaginary difference L_{RI} and magnitude difference L_{Mag} as previously done in other complex spectral mapping studies [44].

$$\begin{aligned} L_{complex} &= L_{RI} + L_{Mag} \\ L_{RI} &= \frac{1}{TF}\sum_{t=1}^{T}\sum_{f=1}^{F}(|\hat{S}_r(t,f) - S_r(t,f)| \\ &\quad + |\hat{S}_i(t,f) - S_i(t,f)|) \\ L_{Mag} &= \frac{1}{TF}\sum_{t=1}^{T}\sum_{f=1}^{F}(|\hat{S}(t,f)| - |S(t,f)|). \end{aligned} \tag{9}$$

A magnitude loss term is added in order to reflect the relative importance of magnitude over phase. Recently, Zhang *et al.* [49] proposed a weighted magnitude-phase loss also in an attempt to emphasize the importance of magnitude estimation.

Finally, our triple-domain loss is the linear combination of the three losses described above. Each component in the triple-domain loss optimizes speech with respect to a different signal representation.

$$L_{triple} = \lambda_1 L_{mask} + \lambda_2 L_{time} + L_{complex}, \tag{10}$$

where λ_1 and λ_2 are the coefficients to balance different value ranges of the three loss terms. We empirically select $\lambda_1 = 5.0$, $\lambda_2 = 1.0$, based on the performance on the validation data.

Xu *et al.* [48] proposed a components loss, consisting of three components for speech preservation, noise suppression, and residual noise quality. Compared with single-component baselines, their experiments show improved and balanced performance. Unlike our triple-domain loss, the components loss is defined in the same STFT domain. Also, our triple-domain loss is different from the triplet loss used in automatic speaker recognition [17], which is defined in terms of a positive sample and a negative sample in order to maximize intra-class similarities and inter-class differences.

IV. EXPERIMENTAL SETTINGS

A. Dataset

The proposed cascade architecture is evaluated on the WSJ0 SI-84 dataset [30], which consists of 7138 utterances from 83 speakers (42 males and 41 females). We select 5428 utterances from 77 speakers to generate the training set. In addition, 20000 noises are randomly chosen from the DNS-Challenge[1] as our training noises, which have a total duration of approximately 55 hours. To generate training mixtures, we randomly cut a segment from the training noises, and then mix it with a randomly picked training utterance at a SNR level that is uniformly sampled from $\{-5, -4, -3, -2, -1, 0\}$ dB. Using this strategy we create a training set with 50000 mixtures. We set aside 150 clean utterances from the training data to create a validation set that is composed of 4000 mixtures. For test purposes, we use 4 challenging noises that are babble (identified as babble1) and factory1 from NOISEX92 [39], and babble (identified as babble2) and cafeteria from an Auditec CD.[2] Test data are generated by mixing these noises with 150 utterances selected from 6 untrained speakers (25 each) at three SNR levels $\{-5, 0, 5\}$ dB.

B. Experimental Setup

All the utterances are sampled at 16 kHz. For STFT operations, we use a 20 ms Hamming window with 50% overlap between adjacent time frames. That is, we use 320-point STFT, which corresponds to a 161-dimensional spectrum. For the frame-level processing in the time module, each utterance is divided into segments of 2048 samples (i.e 128 ms segment), and with an overlap of 1024 samples between consecutive segments.

[1][Online]. Available: https://github.com/microsoft/DNS-Challenge
[2][Online]. Available: http://www.auditec.com

Both causal and non-causal networks are trained with stochastic gradient descent optimization using the Adam optimizer [15]. We train the models for 50 epochs with a batch size of 8 utterances, and set an initial learning rate of 0.001. Utterances that are longer than 8 seconds are chunked to stabilize training, and shorter utterances in a batch are padded with zeros such that all inputs have the same size. Note that during loss calculation, the zero-padded region is ignored. The learning rate is halved if the loss on the validation set has not decreased for 3 consecutive epochs. Gradient clipping with a maximum value 5.0 is applied to avoid gradient explosion.

In our experiments, the performance is assessed by using standard speech enhancement metrics extended short-term objective intelligibility (ESTOI) [14], and perceptual evaluation of speech quality (PESQ) [32]. ESTOI typically has a value range from 0 to 1 and can be interpreted as percent correct, and PESQ has a value range from -0.5 to 4.5. For both metrics, higher values indicate better results.

C. Baselines for Comparison

We compare the proposed cascade architecture with six strong baselines. We first compare with the self-attentive temporal convolutional network (SATCN) [18] that performs multi-stage enhancement in magnitude spectrogram. We adopt the 5-stage configuration that stacks 5 TCNs and incrementally refines the magnitude estimation, resulting in a model with 9.91 M parameters. The second one is the gated CRN (GCRN) [37] that conducts complex spectral mapping. GCRN has a similar structure to CRN, and is also composed of 5 convolutional and deconvolutional blocks for the encoder and the decoder modules, respectively. The major difference is that each convolution is combined with a gated linear unit, and two decoders are used to for estimate real and imaginary parts separately. We follow the configuration described in the original paper that has 9.77 M parameters. The third one is the autoencoder CNN (AECNN) [26], which is a fully convolutional network that operates in the time domain. It is an autoencoder network that is fed with noisy speech segments, and predicts the corresponding clean speech segments. We follow the original description and replicate the network with 6.32 M parameters. For training, we use a segment size of 16384 samples with 50% overlap for the WSJ0 SI-84 dataset. Deep Complex Convolution Recurrent Network (DCCRN) [12] is selected as the fourth baseline, which achieved the first rank in the real-time track of the 2020 Interspeech DNS-Challenge. It is a complex version of CRN, and incorporates complex operations for both CNN and LSTM layers. For comparison, we choose the DCCRN-E configuration described in the paper with around 3.67 M parameters. We also include a two-stage enhancement approach that performs masking and inpainting (M&I) [10] on the noisy magnitude spectrogram. The binary masking and spectrogram inpainting modules are similar, and each is implemented using a residual neural network. In addition, we replace the standard convolutions with partial convolutions in the inpainting module. M&I is a non-causal model that uses 160×160 magnitude spectrograms as the input, and has 20.47 M parameters. Another two-stage baseline for comparison is the Complex spectral mapping based Two-Stage Network (CTSNet) [16]. CTSNet consists of two temporal convolution based modules that progressively enhance noisy speech. The first module estimates the magnitude spectrogram and the second performs complex spectral mapping. The number of intermediate channels is set to 64 to be consistent with the settings in [16], resulting in a model with around 6.55 M parameters.

V. Results, Comparisons and Analyses

A. Evaluation and Comparison Results

In this section, we present evaluation results on the WSJ0 SI-84 dataset, and compare the performance of our cascade architecture with the recent baselines in both causal and non-causal settings. The results are provided in II and III, in terms of ESTOI and PESQ for four challenging nonstationary noises at the SNR levels of -5 dB, 0 dB, and 5 dB. We highlight the best score under each condition by boldface. We observe that all the DNN-based speech enhancement models effectively remove noises in various conditions for untrained speakers and noises. In addition, the proposed neural cascade architecture (NCA) yields the best results in all conditions.

Under the causal settings, the tables show that our NCA model substantially outperforms the time-domain AECNN in both metrics. For example, under -5 dB SNR, we see ESTOI improved by 9.81% and PESQ by 0.18. In addition, SATCN only operates on magnitude spectrograms and performs worse than other baselines even though it performs enhancement in multiple stages. GCRN and DCCRN perform complex spectral mapping in one stage, and CTSNet optimizes the complex spectrogram in multiple stages. Compared with the other baselines, CTSNet shows a significant advantage. Moreover, our model consistently outperforms CTSNet, particularly in ESTOI; for example, on average ESTOI is improved by 3.94%, and PESQ by 0.07 under -5 dB SNR conditions.

We also provide the non-causal enhancement results under the exact same experimental setup. To make the baseline models non-causal, we replace the causal convolutions with non-causal convolutions. Furthermore, all LSTM layers are replaced with bidirectional LSTMs. The non-causal models are denoted as NC-SATCN, BGCRN, BDCCRN, NC-CSTNet and NC-NCA; non-causal AECNN is not included as turning AECNN into a non-causal version is not straightforward. Not surprisingly, there is a substantial performance gap between causal models and their non-causal counterparts because non-causal models utilize future information. With non-causal settings our cascade architecture maintains a consistent performance advantage. In fact, our proposed network provides an even larger performance gain over the best baseline of NC-CTSNET, for example by 5.87% ESTOI and 0.22 PESQ on average in the -5 dB SNR condition.

Fig. 5 illustrates the spectrograms of an example utterance from the test set mixed at -5 dB with the babble2 noise. Fig. 5(a)–(e) show respectively the mixture spectrogram, magnitude-masked spectrogram, time-domain enhanced spectrogram, final enhanced spectrogram, and clean speech spectrogram. The figures demonstrate that the cascade architecture progressively improves the enhancement result.

TABLE II
EVALUATIONS AND COMPARISONS OF DIFFERENT ENHANCEMENT MODELS IN TERMS OF ESTOI(%)

SNR	Causal	-5 dB					0 dB					5 dB				
		Babble1	Factory	Babble2	Cafeteria	Average	Babble1	Factory	Babble2	Cafeteria	Average	Babble1	Factory	Babble2	Cafeteria	Average
Mixture	-	26.40	26.75	25.62	24.25	25.76	39.47	40.94	38.81	37.92	39.29	54.05	56.77	53.89	53.44	54.54
SATCN	✓	44.88	45.62	45.18	42.48	44.54	62.95	63.96	64.02	62.11	63.26	75.78	76.65	76.95	75.28	76.14
GCRN	✓	51.10	52.56	51.40	48.40	50.87	68.72	69.17	69.87	67.05	68.71	80.45	80.68	81.74	79.45	80.58
AECNN	✓	57.88	59.72	48.87	47.81	53.57	71.87	72.21	66.02	64.97	68.77	79.62	79.71	79.12	78.18	79.16
DCCRN	✓	54.72	55.41	54.78	52.39	54.33	71.71	71.59	72.68	70.38	71.59	82.66	82.59	83.89	81.94	82.77
CTSNet	✓	60.06	61.05	60.21	56.43	59.44	75.86	75.10	74.74	74.03	74.93	84.54	83.89	86.44	84.48	84.84
NCA	✓	**63.96**	**63.41**	**65.76**	**60.40**	**63.38**	**78.83**	**77.75**	**80.56**	**76.89**	**78.51**	**86.50**	**85.82**	**87.73**	**85.37**	**86.36**
M&I	✗	36.61	42.78	36.68	37.80	37.72	55.36	59.41	55.79	55.80	56.59	69.45	72.31	70.55	70.25	70.64
NC-SATCN	✗	50.94	50.48	50.50	48.39	50.08	69.05	68.12	68.13	66.58	67.96	79.41	79.60	80.32	78.44	79.44
BGCRN	✗	56.83	58.79	57.34	54.71	56.92	73.52	73.79	75.00	72.31	73.66	83.66	83.64	84.64	82.72	83.67
BDCCRN	✗	57.46	58.79	58.26	55.78	57.57	74.76	74.25	75.65	73.13	74.45	84.50	84.11	85.56	83.54	84.43
NC-CTSNet	✗	63.01	63.42	63.09	60.24	62.44	78.91	77.51	79.91	76.94	78.32	86.19	85.70	87.47	85.48	86.21
NC-NCA	✗	**69.06**	**68.29**	**70.61**	**65.26**	**68.31**	**81.71**	**80.55**	**83.54**	**80.00**	**81.45**	**88.10**	**87.32**	**89.23**	**87.15**	**87.95**

TABLE III
EVALUATIONS AND COMPARISONS OF DIFFERENT ENHANCEMENT MODELS IN TERMS OF PESQ

SNR	Causal	-5 dB					0 dB					5 dB				
		Babble1	Factory	Babble2	Cafeteria	Average	Babble1	Factory	Babble2	Cafeteria	Average	Babble1	Factory	Babble2	Cafeteria	Average
Mixture	-	1.54	1.44	1.55	1.46	1.50	1.83	1.75	1.82	1.77	1.79	2.15	2.10	2.11	2.13	2.12
SATCN	✓	1.64	1.73	1.69	1.61	1.67	2.21	2.33	2.21	2.24	2.25	2.70	2.76	2.72	2.70	2.72
GCRN	✓	1.71	1.95	1.73	1.75	1.79	2.37	2.56	2.43	2.43	2.45	2.94	3.03	2.98	2.93	2.97
AECNN	✓	2.10	2.31	1.82	1.95	2.05	2.68	2.36	2.54	2.43	2.50	2.90	2.92	2.89	2.91	2.91
DCCRN	✓	1.98	2.08	1.93	1.98	1.99	2.49	2.52	2.52	2.49	2.51	2.92	2.89	2.97	2.91	2.92
CTSNet	✓	2.12	2.29	2.13	2.11	2.16	2.74	2.78	2.67	2.78	2.74	3.21	3.18	3.22	3.17	3.20
NCA	✓	**2.15**	**2.35**	**2.22**	**2.20**	**2.23**	**2.81**	**2.88**	**2.91**	**2.83**	**2.86**	**3.27**	**3.26**	**3.34**	**3.23**	**3.28**
M&I	✗	1.77	1.94	1.72	1.84	1.82	2.21	2.36	2.14	2.26	2.24	2.68	2.81	2.61	2.73	2.71
NC-SATCN	✗	1.91	2.00	1.83	1.89	1.91	2.46	2.5	2.56	2.48	2.50	3.05	3.08	3.02	3.03	3.05
BGCRN	✗	2.08	2.31	2.06	2.10	2.14	2.31	2.75	2.82	2.74	2.66	3.20	3.23	3.19	3.18	3.20
BDCCRN	✗	2.05	2.23	2.06	2.11	2.11	2.65	2.68	2.67	2.62	2.66	3.05	3.02	3.08	3.01	3.04
NC-CTSNet	✗	2.19	2.39	2.20	2.23	2.25	2.90	2.91	2.93	2.85	2.90	3.32	3.29	3.33	3.27	3.30
NC-NCA	✗	**2.43**	**2.54**	**2.49**	**2.40**	**2.47**	**3.05**	**3.05**	**3.13**	**3.02**	**3.06**	**3.43**	**3.40**	**3.47**	**3.40**	**3.43**

TABLE IV
EFFECT OF VARIOUS OPTIMIZATION STRATEGIES

	-5 dB SNR		
Model	ESTOI(%)	PESQ	Training time
End-to-end optimization	**63.38**	**2.23**	1.0x
Multi-stage sequential training	60.81	2.11	1.9x
Multi-stage joint training	63.09	**2.23**	2.1x
Only optimizing $L_{complex}$	58.56	2.09	1.0x

TABLE V
ABLATION STUDY ON COMPONENTS OF THE CASCADE ARCHITECTURE

	-5 dB SNR		
Model	ESTOI%	PESQ	params (M)
NCA	**63.38**	**2.23**	12.9
−DC	59.82	2.03	12.7
−1x1Conv Skip	62.40	2.21	12.8
−Noisy Input	62.06	2.18	12.7

The spectrogram enhanced by our model effectively removes the background noise even at −5 dB, and is close to that of the clean speech. We provide enhanced speech samples at https://whmrtm.github.io/NCA_demo.html.

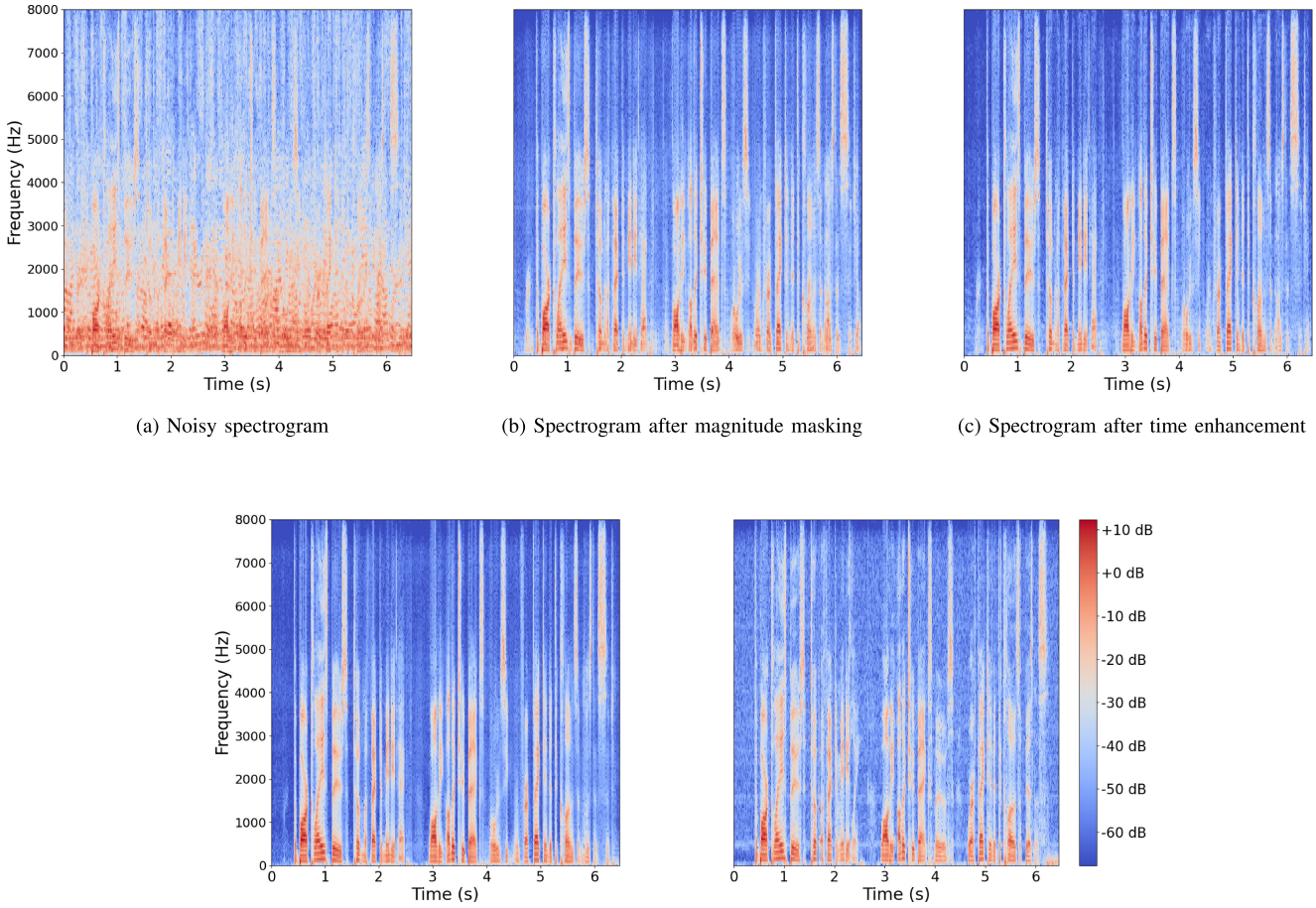
B. *Effects of Optimization Strategies*

We further evaluate the proposed NCA with various optimization strategies. Table IV compares the enhancement performance and training time of several strategies under −5 dB SNR. The first row of the table is the result of end-to-end training, where all modules are updated simultaneously. There are two reasonable multi-stage training strategies. One is the sequential training strategy. First, we train CRN-Mask to predict the IRM. Next, with the first module frozen, we train UNet-Time by taking the noisy input and the output from the frozen CRN-Mask. Finally, we train the last module CRN-Complex with the previous modules frozen, and take as input the prediction from the second module and the noisy input. The other strategy is similar, but without freezing previous modules. Specifically, for the second module, we jointly train the pre-trained CRN-Mask and UNet-Time, but use a smaller learning rate for fine-tuning the first module. In the last step, the last two modules are jointly trained, with the first module frozen. Similarly, the second module is fine-tuned with a smaller learning rate. Our experiments show that end-to-end optimization and multi-stage joint training have the best performance. Sequential training degrades the enhancement performance; for example, at −5 dB SNR, ESTOI is decreased by 2.58% and PESQ by 0.12. The end-to-end optimization strategy would be preferable, because it is conceptually simpler and does not require extra hyperparameters. In addition, this strategy saves training effort. Compared with the end-to-end training, the joint training strategy reaches similar performance using more than twice training time.

Table IV also compares our model trained with only the complex loss $L_{complex}$ measured at the last module (last row) and the cross-domain loss measured in all three modules with the same cascade architecture. Only optimizing $L_{complex}$ degrades the results considerably compared to optimizing the proposed triple-domain loss, as the first two modules are not constrained by their respective objectives. This comparison demonstrates that the strong performance of the cascade architecture is not entirely due to the neural network structure, and the training strategy with the triple-domain loss is also a contributing factor.

C. *Ablation Study*

To investigate the contributions of the introduced techniques, we conduct an ablation study using the causal cascade architecture at −5 dB SNR. As shown in Table V, we evaluate several



(a) Noisy spectrogram (b) Spectrogram after magnitude masking (c) Spectrogram after time enhancement

(d) Spectrogram enhanced after complex mapping (e) Clean spectrogram

Fig. 5. Spectrograms of example speech enhancement: (a) Noisy speech mixture, (b) Speech enhanced by the first module, (c) Speech enhanced by the second module, (d) Speech enhanced by the proposed architecture, and (e) Clean speech.

variants for comparisons under by taking the average results of all test noises. In the first variant, denoted as − DC, we remove dense connections in the complex module. In the second one, denoted as − 1×1 Conv Skip, we replace the pointwise convolutional skip connections with naive concatenations. The last one, denoted as − Noisy Input, removes the noisy input for the second and third modules. The experimental results in Table V show that these variants underperform the original cascade architecture. Among these components, dense connections play a significant role, contributing 3.56% ESTOI and 0.20 PESQ. In addition, it is important that each module has access to the noisy input.

In Table VI, we present evaluation results under the causal setting to analyze the contribution of each part. We have investigated using only a single module, and 2-module and 3-module combinations. In order to have a fair comparison, we adjust the number of convolutional channels such that all variants in the table have comparable numbers of parameters. As shown in the table, combining cross-domain modules is advantageous over using single modules. Among two-module models, coupling CRN-Mask and UNet-Time performs slightly better than the others in terms of PESQ. Coupling UNet-Time and CRN-Complex, however, exhibits better ESTOI scores. Furthermore, among reasonable combinations of three modules, our proposed model has the best overall performance.

TABLE VI
EXPERIMENTAL COMPARISONS OF MODULE COMBINATIONS

	-5 dB SNR		
Model	ESTOI%	PESQ	params (M)
Proposed	**63.38**	**2.23**	12.9
CRN-Mask alone	47.24	1.84	12.4
UNet-Time alone	54.96	1.93	12.4
CRN-Complex alone	53.50	2.05	12.9
CRN-Mask + CRN-Complex	59.31	2.01	13.3
CRN-Mask + UNet-Time	59.69	2.22	13.1
UNet-Time + CRN-Complex	60.76	2.14	13.4
CRN-Mask + CRN-Complex + UNet-Time	61.53	**2.24**	12.9
CRN-Complex + CRN-Mask + UNet-Time	54.94	2.08	12.9
UNet-Time + CRN-Mask + CRN-Complex	60.93	2.12	12.9

D. Comparison of Model Complexities

Table VII lists the model size and multiply-accumulate operations (MACs) of the proposed cascade architecture and the other baselines in the causal settings. The model size is the number of trainable parameters within a model, and the computational complexity is calculated by taking the average of enhancing 50 test utterances using an open-source package.[3] As shown in the table, our model achieves superior performance with reasonable computational complexity. The NCA model has about 12.87

[3][Online]. Available: https://github.com/Lyken17/pytorch-OpCounter

TABLE VII
NUMBER OF TRAINABLE PARAMETERS AND MACS FOR DIFFERENT ENHANCEMENT MODELS, WHERE M INDICATES MILLION

Model	Parameters	MACs
CRN-Mask	4.58 M	28.89 M
UNet-Time	3.47 M	31.88 M
CRN-Complex	4.82 M	39.26 M
NCA	12.87 M	100.03 M
SATCN	9.91 M	64.94 M
GCRN	9.77 M	25.07 M
DCCRN	3.67 M	155.33 M
AECNN	6.32 M	186.22 M
CTSNet	6.55 M	84.70 M
M&I	20.47 M	207.27 M

million parameters, and about 100 million MACs. Additionally, the first three rows display the numbers of parameters and MACs of each module within NCA, and they are reasonably comparable. In general, DCCRN has the smallest number of parameters, and GCRN is the fastest model.

VI. CONCLUSION

In this study, we have proposed a novel cascade architecture for monaural speech enhancement. The key idea is to leverage the strengths of time and frequency domain approaches by progressively enhancing noisy speech in different domains of speech representation. In the cascade architecture, the first module estimates the IRM from the noisy input. The second module performs time-domain enhancement on the output of the first module and the noisy input. The last module further refines the enhanced speech by performing complex spectral mapping on the output of the second module and the noisy input. Experimental results demonstrate the superiority of the neural cascade architecture trained with the triple-domain loss.

We have explored various combinations of the modules, confirming that the proposed cascade design works the best. Also, we optimize all the modules simultaneously, avoiding the complexities of pre-training and fine-tuning. For future research, we plan to make our model practical by reducing the network complexity and inference time (see [35]). Additionally, we plan to extend our cascade architecture to speech enhancement in multi-channel and reverberant environments.

ACKNOWLEDGMENT

The authors would like to thank Ashutosh Pandey and Ke Tan for insightful discussions and assistance in model comparisons.

REFERENCES

[1] F. Bahmaninezhad *et al.*, "A comprehensive study of speech separation: Spectrogram vs waveform separation," in *Proc. INTERSPEECH*, 2019, pp. 4574–4578.
[2] J. Chen and D. L. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *J. Acoustical Soc. Amer.*, vol. 141, pp. 4705–4714, 2017.
[3] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex U-Net," in *Proc. Int. Conf. Learn. Representations*, 2018.
[4] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5024–5028.
[5] C. Fan, J. Tao, B. Liu, J. Yi, Z. Wen, and X. Liu, "End-to-end post-filter for speech separation with deep attention fusion features," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1303–1314, 2020, doi: 10.1109/TASLP.2020.2982029.
[6] S.-W. Fu, T.-y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *Proc. Int. Workshop Mach. Learn. Signal Proc.*, 2017, pp. 1–6.
[7] S.-W. Fu, Y. Tsao, and X. Lu, "SNR-aware convolutional neural network modeling for speech enhancement," in *Proc. INTERSPEECH*, 2016, pp. 3768–3772.
[8] F. Gao, L. Wu, L. Zhao, T. Qin, X. Cheng, and T.-Y. Liu, "Efficient sequence learning with group recurrent networks," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2018, pp. 799–808.
[9] K. Han, Y. Wang, and D. L. Wang, "Learning spectral mapping for speech dereverberation," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 4628–4632.
[10] X. Hao *et al.*, "Masking and inpainting: A two-stage speech enhancement approach for low SNR and non-stationary noise," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6959–6963.
[11] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.
[12] Y. Hu *et al.*, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. INTERSPEECH*, 2020, pp. 2482–2486.
[13] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 4700–4708.
[14] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.
[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Int. Conf. Learn. Representations*, 2015.
[16] A. Li, C. Zheng, R. Peng, and X. Li, "Two heads are better than one: A two-stage approach for monaural noise reduction in the complex domain," 2020, *arXiv:2011.01561*.
[17] C. Li *et al.*, "Deep speaker: An end-to-end neural speaker embedding system," 2017, *arXiv:1705.02304*.
[18] J. Lin, A. J. van Wijngaarden, K.-C. Wang, and M. C. Smith, "Speech enhancement using multi-stage self-attentive temporal convolutional networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3440–3450, 2021.
[19] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC press, 2013.
[20] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 46–50.
[21] C. Macartney and T. Weyde, "Improved speech enhancement with the Wave-U-Net," 2018, *arXiv:1811.11307*.
[22] P. Mowlaee, R. Saeidi, and R. Martin, "Phase estimation for signal reconstruction in single-channel source separation," in *Proc. INTERSPEECH*, 2012, pp. 1548–1551.
[23] Z. Ouyang, H. Yu, W.-P. Zhu, and B. Champagne, "A fully convolutional neural network for complex spectrogram processing in speech enhancement," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 5756–5760.
[24] A. Pandey and D. L. Wang, "Dense CNN with self-attention for time-domain speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1270–1279, 2021, doi: 10.1109/TASLP.2021.3064421.
[25] A. Pandey and D. L. Wang, "Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6629–6633.
[26] A. Pandey and D. L. Wang, "A new framework for CNN-based speech enhancement in the time domain," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 7, pp. 1179–1188, Jul. 2019.
[27] A. Pandey and D. L. Wang, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6875–6879.

[28] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," in *Proc. INTERSPEECH*, 2016, pp. 1993–1997.
[29] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," in *Proc. INTERSPEECH*, 2017, pp. 3642–3646.
[30] D. B. Paul and J. Baker, "The design for the wall street journal-based CSR corpus," in *Proc. a Workshop Speech Natural Lang.*, 1992, pp. 23–26.
[31] H. Phan *et al.*, "Improving GANs for speech enhancement," *IEEE Signal Process. Lett.*, vol. 27, pp. 1700–1704, 2020, doi: 10.1109/LSP.2020.3025020.
[32] 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," *Proc. Int. Conf. Acoust., Speech Signal Process.*, vol. 2, pp. 749–752, 2001.
[33] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.
[34] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *Proc. Hands-Free Speech Commun. Microphone Arrays*, 2017, pp. 136–140.
[35] K. Tan and D. L. Wang, "Towards model compression for deep learning based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1785–1794, 2021, doi: 10.1109/TASLP.2021.3082282.
[36] K. Tan and D. L. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Proc. INTERSPEECH*, 2018, pp. 3229–3233.
[37] K. Tan and D. L. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 380–390, 2019, doi: 10.1109/TASLP.2019.2955276.
[38] E. Tzinis, S. Venkataramani, Z. Wang, C. Subakan, and P. Smaragdis, "Two-step sound source separation: Training on learned latent targets," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 31–35.
[39] 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 Commun.*, vol. 12, pp. 247–251, 1993.
[40] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*. Boston, MA, USA: Springer, 2005, pp. 181–197.
[41] D. L. Wang and G. J. Brown, Eds, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ, USA: Wiley, 2006.
[42] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.
[43] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.
[44] Z.-Q. Wang, P. Wang, and D. L. Wang, "Complex spectral mapping for single-and multi-channel speech enhancement and robust ASR," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1778–1787, 2020, doi: 10.1109/TASLP.2020.2998279.
[45] F. Weninger *et al.*, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, 2015, pp. 91–99.
[46] D. S. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 483–492, Mar. 2016.
[47] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, Jan. 2014.
[48] Z. Xu, S. Elshamy, Z. Zhao, and T. Fingscheidt, "Components loss for neural networks in mask-based speech enhancement," *EURASIP J. Audio, Speech, Music Process.*, vol. 2021, pp. 1–20, 2021.
[49] J. Zhang, M. D. Plumbley, and W. Wang, "Weighted magnitude-phase loss for speech dereverberation," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 5794–5798.
[50] Y. Zhao, Z.-Q. Wang, and D. L. Wang, "Two-stage deep learning for noisy-reverberant speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 53–62, Jan. 2018.

Heming Wang (Student Member, IEEE) received his Bachelor degree in Physics in 2016, and the M.S. degree in Applied Mathematics in 2018 from University of Waterloo, Ontario, Canada. He is currently working toward the Ph. D. degree at the Ohio State University. His research interests lie in speech enhancement and deep learning.

DeLiang Wang (Fellow, IEEE) Author photograph and biography not available at the time of publication.