IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 25, 2020, accepted October 4, 2020, date of publication October 7, 2020, date of current version October 19, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3029417

Improved Relativistic Cycle-Consistent GAN With Dilated Residual Network and Multi-Attention for Speech Enhancement

YUTIAN WANG[1], (Member, IEEE), GUOCHEN YU[1], JINGLING WANG[1], HUI WANG[2], (Member, IEEE), AND QIN ZHANG[1,2]

[1]Key Laboratory of Media Audio and Video, Ministry of Education, Communication University of China, Beijing 100024, China
[2]State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China

Corresponding author: Hui Wang (hwang@cuc.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61631016 and Grant 61501410, and in part by the Fundamental Research Funds for the Central Universities under Grant 3132018XNG1805.

ABSTRACT Generative adversarial networks (GANs) have been increasingly used as feature mapping functions in speech enhancement, in which the noisy speech features are transformed to the clean ones through the generators. This article proposes a novel speech enhancement model based on a cycle-consistent relativistic GAN with Dilated Residual Networks and a Multi-attention mechanism. Using the adversarial loss, improved cycle-consistency losses, and an identity-mapping loss, a noisy-to-clean generator G and an inverse clean-to-noisy generator F simultaneously learn the forward and backward mappings between the source and target domains. To guarantee the stability of the training process, we replace vanilla GAN loss with relativistic average GAN loss and use spectral normalization in discriminators so that they conform to Lipschitz continuity. Furthermore, we employ two attention-based components as multi-attention mechanism to reduce importing signal distortion: attention U-net gates and dilated residual self-attention blocks. By employing these components, our proposed generators can capture long-term inner dependencies between elements of speech features and further preserve linguistic information. Experimental results on a public dataset indicate that the proposed model achieves state-of-the-art speech enhancement performance, especially in reducing speech distortion and improving signal overall quality. Compared with the representative GAN-based approaches, the proposed method significantly achieves the best performance in terms of STOI, CSIG, COVL, and CBAK objective metrics. Moreover, we demonstrate the contribution of each proposed component including relativistic average loss, attention U-net gate, self-attention layers, spectral normalization, and dilation operation by ten comparison systems.

INDEX TERMS Speech enhancement, cycle-consistent GAN, relativistic average loss, multi-attention, dilated residual network, U-net.

I. INTRODUCTION

Speech enhancement removes additive noisy interferences from noisy speech signal while preserving the intelligibility of the original clean speech. Speech enhancement approaches not only improve the speech intelligibility and quality but also work as a preprocessor for other downstream speech applications, such as robust automatic speech recognition [1], speaker identification [2], and hearing aids [3].

Past speech enhancement methods fall into two categories: statistical-based approaches and data-driven approaches. Statistical-based methods are based on particular probabilistic models of noisy speech [4], [5], such as spectral subtraction methods [6], Wiener filtering [7], and minimum mean-square error (MMSE) of the spectral amplitude [8]. Recently, data-driven approaches use deep neural networks (DNNs) to estimate the ideal ratio mask (IRM) or ideal binary mask (IBM) and have demonstrated significant performance improvement over the conventional statistical-based methods [9], [10]. However, these methods have the presumption that the scale of the masked signal is the same

The associate editor coordinating the review of this manuscript and approving it for publication was Lefei Zhang.

as the source speech and the noise is strictly additive, which are uncommon in the real environment. To address these limitations, some approaches (e.g., denoising auto-encoders [11]) train a direct mapping network from the noisy input features to the enhanced ones. A recent breakthrough comes from the application of generative adversarial networks (GANs) [12] as the feature mapping networks. GAN consists of a generator network (G) and a discriminator network (D) that play a min-max game between each other and achieves impressive results in the computer vision field. By using the adversarial training, D tries to distinguish the generated data by G from the real data, on the other hand, G forces the generated data to be indistinguishable from the real one. In the speech area, GANs have been widely used in speech generation [13], speech enhancement [14]–[18], voice conversion [19], and acoustic model adaptation [20].

Conventional GAN-based models achieve state-of-the-art performance for speech enhancement, but for improving the generalization of the models, they always require a large quantity of parallel datasets. However, it is difficult to obtain parallel clean speech and noise data from a real scenario. To solve this problem, using cycle-consistent GAN (CycleGAN) [21] for speech enhancement becomes a good strategy, preserving the speech structure and improving speech enhancement performance. Inspired by recent studies on CycleGAN-based approaches for speech processing [22]–[24], we propose a relativistic-loss cycle-consistent GAN with multi-attention and dilated residual network (DRN) for single-channel speech enhancement. This model contains a noisy-to-clean generator G and an inverse clean-to-noisy generator F, which transforms the noisy features into the enhanced ones and vice versa. In this model, the forward noisy-clean-noisy cycle and backward clean-noisy-clean cycle are jointly trained with the adversarial loss, a cycle-consistency loss, cycle-adversarial losses, and an identity-mapping loss. Subsequently, to address the difficulty of finding a Nash equilibrium of a min-max game between generators and discriminators in GAN training, the relativistic average least-square loss [25] is adopted to substitute conventional cross-entropy loss or least-square loss [26]. Spectral normalization [27] is also applied to stabilize GAN training and generate better samples.

To further preserve linguistic information and capture contextual relationship in enhancing process, a multi-attention mechanism is employed in U-shape generators, which consist of encoding layers, transformation blocks, and decoding layers. Attention mechanism can compute the long-range relative dependencies among elements in sequences [28], which has been widely used both in the computer vision field and speech area [29], [30]. The proposed multi-attention uses attention mechanism in two different ways: attention gates in U-net [31] encoding-decoding layers (AU gate) and self-attention [32] in dilated residual networks [33] (DRN-SA block). With the AU gate, the encoding-decoding layers yield multi-scale features that are better for making predictions by connecting the encoding layers to homologous decoding layers, selectively focusing on salient speech linguistic information in feature enhancing procedure. The DRN-SA block models the long-term relative dependencies between different positions of compressed feature maps and also inherits the advantages of the residual network, better maintaining the contextual temporal structure of features through residual connections. Moreover, the DRN-SA block can make up the reduction of the receptive field with dilation operation, demonstrating a strong ability in modeling local contextual relationships. The objective evaluation results on a public dataset [34] demonstrate that the proposed method obtains state-of-the-art performance in terms of perceptual evaluation, speech intelligibility, overall signal quality, and speech distortion. Moreover, the ablation study indicates each component, including the relativistic loss, spectral normalization, AU gates, and DRN-SA blocks, obviously improves speech enhancement performance.

The rest of this article is organized as follows: In Section II, the whole network training procedure of our speech enhancement approach is illustrated, followed by a description of the detailed network architectures of proposed generators and discriminators in Section III. Then, the experimental setup and preprocessing are presented in Section IV, before we show the objective evaluation and results in Section V. Finally, we conclude the paper and suggest future work in Section VI.

II. CYCLE-CONSISTENT RELATIVISTIC GAN FOR SE

Cycle-consistent GANs learn the forward and backward mappings between source features $x \in \mathrm{X}$ and target features $y \in \mathrm{Y}$ with generators G and F. For the speech enhancement task, G learns a noisy-to-clean mapping and F learns an inverse clean-to-noisy mapping which reconstructs the noisy features from the enhanced ones. Discriminators D are trained to classify samples from the target speech features as real and the generated speech features from generators as fake. The whole architecture of the proposed algorithm is shown in Figure 1.

As illustrated in Figure 2, the forward noisy-clean-noisy cycle and backward clean-noisy-clean cycle are jointly trained to constrain G and F to be cycle-consistent. Our model uses the following losses to jointly optimize the enhancement process, namely relativistic adversarial loss, improved cycle-consistency loss, and identity mapping loss.

A. RELATIVISTIC ADVERSARIAL LOSS

For the noisy-to-clean mapping, the adversarial loss is used to make the enhanced speech features $G_{X\rightarrow Y}(x)$ indistinguishable from the clean ones y. Simultaneously, we introduce a similar adversarial loss for the inverse clean-to-noisy mapping.

To stabilize the training process and improve the quality of the generated features, we introduce relativistic GAN loss to substitute the conventional GAN loss. As discussed in [25], the original GAN loss misses a key property that as the probability of generated data being real (i.e., $D_Y(G_{X\rightarrow Y}(x))$)

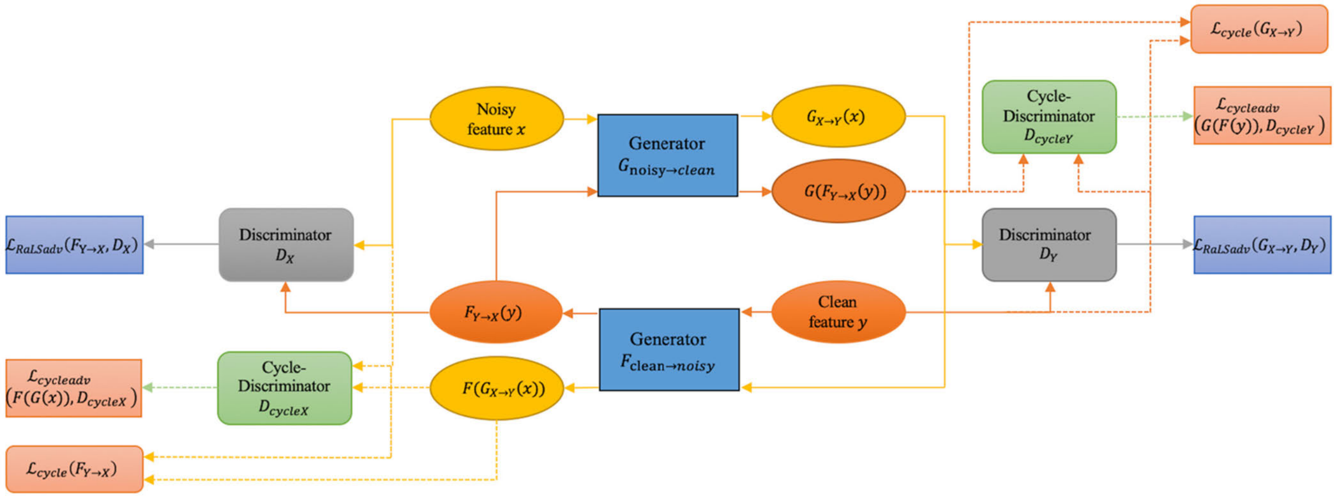

$\mathcal{L}_{cycle}(G_{X\to Y})$
$\mathcal{L}_{cycleadv}(G(F(y)), D_{cycleY})$
Cycle-Discriminator D_{cycleY}
Noisy feature x
Generator $G_{noisy\to clean}$
$G_{X\to Y}(x)$
$G(F_{Y\to X}(y))$
$\mathcal{L}_{RaLSadv}(F_{Y\to X}, D_X)$
Discriminator D_X
Discriminator D_Y
$\mathcal{L}_{RaLSadv}(G_{X\to Y}, D_Y)$
$F_{Y\to X}(y)$
Generator $F_{clean\to noisy}$
Clean feature y
$\mathcal{L}_{cycleadv}(F(G(x)), D_{cycleX})$
Cycle-Discriminator D_{cycleX}
$F(G_{X\to Y}(x))$
$\mathcal{L}_{cycle}(F_{Y\to X})$

FIGURE 1. The tent Relativistic GAN for speech enhancement. Forward noisy-clean-noisy cycle and backward clean-noisy-clean cycle are shown in yellow and red lines, respectively. Generators $G_{noisy\to clean}(x)$ and $F_{clean\to noisy}(y)$ learn the mapping functions between noisy features x and clean features y. Discriminators D_X and D_Y measure the relativistic adversarial loss $\mathcal{L}_{RaLSadv}$, while two cycle discriminators D_{cycleX} and D_{cycleY} calculate the cycle-adversarial loss $\mathcal{L}_{cycleadv}$. Conventional cycle-consistency loss $\mathcal{L}_{cycle}(G_{X\to Y})$ and $\mathcal{L}_{cycle}(G_{X\to Y})$ are jointly optimized to force the mapping functions to be cycle-consistent.

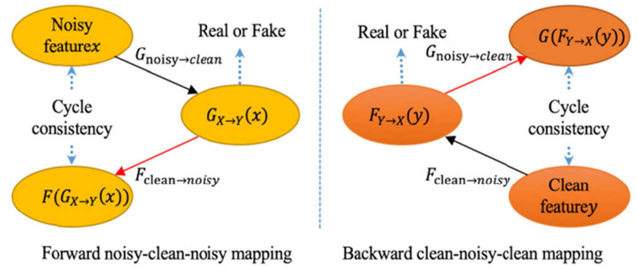


FIGURE 2. The training procedure for Cycle-GAN. Forward noisy-clean-noisy cycle maps the noisy features x to the clean ones and back again, while backward clean-noisy-clean cycle maps the clean features y to the noisy ones $F_{Y\to X}(y)$ and reconstruct back to the clean ones $G(F_{Y\to X}(y))$.

increases, the real data (i.e., $D_Y(y)$) should be simultaneously more difficult to be distinguished as real. To encode this information, the relativistic discriminator estimates a relativistic probability $D_{rel}(x_r, x_g)$, which is defined as,

$$D_{rel}(x_r, x_g) = D(x_r) - D(x_g) \tag{1}$$

where x_r is from the real data and x_g is from the generated data. This probability denotes how much more likely that the given features are real than the generated ones. In contrast, when optimizing generators, this model focuses on another exactly opposite relativistic probability $D'_{rel}(x_r, x_g)$ as,

$$D'_{rel}(x_r, x_g) = D(x_g) - D(x_r) \tag{2}$$

In this article, we adopt the relativistic average least-square GAN (RaLSGAN) losses as our adversarial losses, which are defined as follow:

$$\begin{aligned}\mathcal{L}_{RaLSadv}(D_Y) &= \mathbb{E}_{y\sim P_Y(y)}\left[(D_Y(y) - \mathbb{E}_{x\sim P_X(x)}D_Y(G_{X\to Y}(x)) - 1)^2\right]\\ &\quad + \mathbb{E}_{x\sim P_X(x)}[(D_Y(G_{X\to Y}(x)) - \mathbb{E}_{y\sim P_Y(y)}D_Y(y) + 1)^2]\end{aligned} \tag{3}$$

$$\begin{aligned}\mathcal{L}_{RaLSadv}(G_{X\to Y}) &= \mathbb{E}_{x\sim P_X(x)}\left[(D_Y(G_{X\to Y}(x)) - \mathbb{E}_{y\sim P_Y(y)}D_Y(y) - 1)^2\right]\\ &\quad + \mathbb{E}_{y\sim P_Y(y)}[(D_Y(y) - \mathbb{E}_{x\sim P_X(x)}D_Y(G_{X\to Y}(x)) + 1)^2]\end{aligned} \tag{4}$$

where $\mathcal{L}_{RaLSadv}(D_Y)$ indicates the adversarial loss of discriminator D_Y, and $\mathcal{L}_{RaLSadv}(G_{X\to Y})$ indicates the adversarial loss of noisy-to-clean generator $G_{X\to Y}$. In the above equations, the generator $G_{X\to Y}$ tries to generate enhanced features that can deceive the discriminator D_Y, and D_Y attempts to find the best decision boundary between the clean features and enhanced ones. Similarly, we impose two relativistic average adversarial losses $\mathcal{L}_{RLSadv}(D_X)$ and $\mathcal{L}_{RLSadv}(F_{Y\to X})$ for the inverse noisy-to-clean mapping.

B. IMPROVED CYCLE-CONSISTENCY LOSS

Using adversarial training, G can map noisy features to any random permutation of the clean features, so any learned mapping functions can produce an output distribution that matches the target distribution. In other words, the adversarial loss alone cannot guarantee the contextual information of features x and enhanced features $G_{X\to Y}(x)$ are cycle-consistent. Therefore, as introduced in [21], the cycle-consistency is enforced to preserve speech context integrity by minimizing the cycle-consistency loss as:

$$\begin{aligned}\mathcal{L}_{cycle}(G_{X\to Y}, F_{Y\to X}) &= \mathbb{E}_{x\sim P_X(x)}\left[\| F_{Y\to X}(G_{X\to Y}(x)) - x \|_1\right]\\ &\quad + \mathbb{E}_{y\sim P_Y(y)}[\| G_{X\to Y}(F_{Y\to X}(y)) - y \|_1]\end{aligned} \tag{5}$$

where $\| \, . \, \|_1$ indicates the L1 reconstruction error.

Besides, two additional discriminators D_{cycleX} and D_{cycleY} are introduced to avoid over-smoothing caused by statistical averaging in traditional cycle-consistency loss (e.g., L1 and L2 distances). We propose two novel cycle-adversarial losses $\mathcal{L}_{cycleadv}\left(F_{Y\to X}, D_{cycleX}\right)$ and $\mathcal{L}_{cycleadv}\left(G_{X\to Y}, \mathrm{D}_{cycleY}\right)$ as:

$$\begin{aligned}&\mathcal{L}_{cycleadv}\left(F_{Y\to X}, D_{cycleX}\right)\\&\quad=\mathbb{E}_{x\sim P_X(x)}\left[\log D_{cycleX}\left(x\right)\right]\\&\qquad+\mathbb{E}_{x\sim P_X(x)}\left[log(1-D_{cycleX}(F_{Y\to X}(G_{X\to Y}(x))))\right] \qquad (6)\\&\mathcal{L}_{cycleadv}\left(G_{X\to Y}, D_{cycleY}\right)\\&\quad=\mathbb{E}_{y\sim P_Y(y)}\left[\log D_{cycleY}\left(y\right)\right]\\&\qquad+\mathbb{E}_{y\sim P_Y(y)}\left[log(1-D_{cycleY}(G_{X\to Y}(F_{Y\to X}(y))))\right] \qquad (7)\end{aligned}$$

C. IDENTITY-MAPPING LOSS

We regularize Generators G and F to be close to identity mappings by minimizing identity-mapping loss as in [21]. This loss preserves the compositions ((i.e., linguistic information) of the input source features and the target ones [22], [23], and helps the generators better map the target distribution.

$$\begin{aligned}&\mathcal{L}_{identity}\left(G_{X\to Y}, F_{Y\to X}\right)\\&\quad=\mathbb{E}_{x\sim P_X(x)}\left[\parallel F_{Y\to X}(x)-x\parallel_1\right]\\&\qquad+\mathbb{E}_{y\sim P_Y(y)}\left[\parallel G_{X\to Y}(y)-y\parallel_1\right] \qquad (8)\end{aligned}$$

where real speech features of the target domain (i.e., x and y) are provided as the input to the generators (i.e., $F_{Y\to X}$ and $G_{X\to Y}$), respectively.

The full loss function is written as follows, in which includes above two adversarial losses, a cycle-consistency loss, two cycle-adversarial losses, and an identity-mapping loss:

$$\begin{aligned}\mathcal{L}_{Full}&=\mathcal{L}_{RaLSadv}\left(G_{X\to Y}, D_Y\right)+\mathcal{L}_{RaLSadv}\left(F_{Y\to X}, D_X\right)\\&\quad+\lambda_{cycle}\mathcal{L}_{cycle}\left(G_{X\to Y}, F_{Y\to X}\right)\\&\quad+\mathcal{L}_{cycleadv}\left(G_{X\to Y}, D_{cycleY}\right)\\&\quad+\mathcal{L}_{cycleadv}\left(F_{Y\to X}, D_{cycleY}\right)\\&\quad+\lambda_{id}\mathcal{L}_{identity}\left(G_{X\to Y}, F_{Y\to X}\right) \qquad (9)\end{aligned}$$

where λ_{cycle} and λ_{id} are tunable hyper-parameters.

III. ARCHITECTURE

As mentioned above, two generators, $G_{X\to Y}$ and $F_{Y\to X}$, and four discriminators, D_X, D_Y, D_{cycleX}, and D_{cycleY}, are together employed in our speech enhancement task. For generators, we introduce a multi-attention mechanism and dilated residual networks for speech feature transformation to capture contextual relative dependencies and enlarge receptive field, respectively. Moreover, gated linear units [35] are applied in convolutional layers to model speech sequential structure. As for discriminators, we use spectral normalization [27] in discriminator to limit the weights' numerical ranges and avoid vanishing or exploding gradients

A. IMPROVED TECHNIQUES FOR OUR MODEL

1) GATED LINEAR UNITS

RNN-based approaches are effective to model the sequential and hierarchical structures of speech, but they are computational demanding due to its difficulty of parallelism. Instead, we apply gated CNNs [35] in our architecture, which achieves state-of-the-art performance in language modeling. Similar to the gating mechanisms in long short-term memories (LSTM) and gated recurrent units (GRU), gated CNNs with gated linear units (GLUs) as activation function can selectively propagate information, depending on the previous layers' states. Also, GLUs can alleviate the gradient vanishing problem by providing a linear path for the gradient propagation while simultaneously keeping nonlinear capabilities through a sigmoid gate. The $i+1^{th}$ layer output H_{i+1} of GLU is calculated by the i^{th} layer interval features H_i as,

$$H_{i+1}=\left(H_i*W_i+b_i\right)\otimes\sigma\left(H_i*V_i+c_i\right) \qquad (10)$$

where $\otimes$ denotes the element-wise product, $*$ denotes the convolution operation, and σ indicates the sigmoid activation function. W_i and V_i are convolutional filters with biases b_i and c_i, respectively. The value of the sigmoid function ranges from 0 to 1, by which the network can learn to focus on corresponding speech features and ignore the unrelated ones.

2) IMPROVED NORMALIZATION

We employ instance normalization and spectral normalization in generators and discriminators, respectively. Instance normalization (IN) [36] is successfully used in image stylization and generation, and achieves better performance with less computational cost than batch normalization (BN) or virtual batch normalization (VBN). IN layer applies mean-variance normalization to every channel and single instance instead of a whole batch. Motivated by this, we propose to use instance normalization in generator architecture, reducing the computational cost and improving the enhancement performance of the model.

Spectral normalization (SN) is a novel weight normalization method and can stabilize the training of discriminators [27]. SN constrains the Lipschitz constant of the discriminator by limiting the spectral norm of the weight matrix W in each layer, such that it satisfies the Lipschitz constraint $\sigma\left(W\right)=1$:

$$W_{SN}=W/\sigma\left(W\right) \qquad (11)$$

where $\sigma\left(W\right)$ is the spectral norm of W.

Compared with other normalization techniques for GANs (e.g., weight normalization, weight clipping, and gradient penalty), SN does not require intensive tuning of Lipschitz constant, since it is the only hyper-parameter to be tuned. Moreover, the implementation is simple and the computational complexity is also relatively small.

3) MULTI-ATTENTION MECHANISM

Multi-attention mechanism models the relative dependencies between elements in feature maps and further preserve source

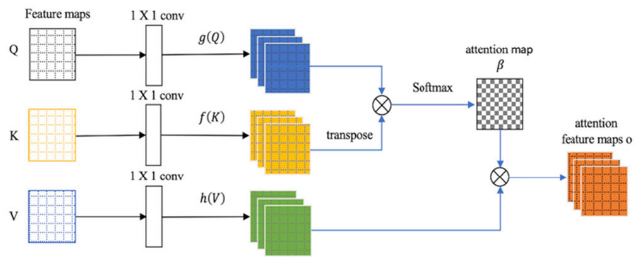


FIGURE 3. The parallel structure of attention mechanism. The ⊗ denotes matrix multiplication. The softmax operation is performed on each row.

linguistic information, considerably enhancing the salient speech information importing in the feature learning procedure. The proposed multi-attention consists of two attention mechanisms, attention U-net gates (AU gates) and dilated residual self-attention blocks (DRN-SA blocks), in encoding-decoding layers and transformation block, respectively.

As shown in Figure 3, attention mechanism maps an input query and a set of key-value pairs to an attention output that is computed as a weighted sum of the values [28]. The weight assigned to each value is calculated by a compatibility function using the input query and the corresponding key. We compute the attention function on a set of queries, keys, and values simultaneously, packing them together into feature maps $Q[[space]] \in \mathbb{R}^{C\times N}$, $K \in \mathbb{R}^{C\times N}$ and $V \in \mathbb{R}^{C\times N}$, respectively. Here, C is the number of channels and N is the number of features. Firstly, the feature maps Q and K are projected into two feature spaces g and f by 1×1 convolutional layers to calculate attention map β s,

$$\beta_{j,i} = softmax\left((W_f K_i)^T (W_g Q_j)\right) \tag{12}$$

where $W_f \in \mathbb{R}^{C\times C}$ and $W_g \in \mathbb{R}^{C\times C}$ are weight matrices, which are learned by 1×1 convolutions; $\beta_{j,i}$ denotes the extent to the long-term dependency between the i^{th} position in feature map K and the j^{th} position in feature map Q.

Then, we compute the output of the attention layer $o = (o_1, o_2, \cdots, o_j, \cdots, o_N) \in \mathbb{R}^{C\times N}$ as,

$$o_j = \sum\nolimits_{i=1}^{N} \beta_{j,i}(W_h V_i) \tag{13}$$

where $W_h \in \mathbb{R}^{C\times C}$ is a weighted matrix of 1×1 CNN layers.

The final output y is the weighted sum of the attention output and the input Q,

$$y = \lambda o + Q \tag{14}$$

where λ is a learnable scalar coefficient and initialized as 0.

B. 2-1-2D ATTENTION-BASED GENERATOR

We design our network based on recent researches on voice conversion and speech modeling [22], [24]. The generator contains three components: the encoding layers, homologous decoding layers, and a transformation block (Figure 4).

As described in [24], two-dimensional (2-D) CNNs are more suitable for feature-transformation while preserving the original speech structures, because it restricts the transformation process focus more on the local region. Inspired by this, we use 2-D convolution in the encoding and decoding layers. In contrast, one-dimensional (1-D) CNNs are better to capture the relationship among the overall features with the feature dimension. 1-D CNNs in residual blocks can mitigate the loss of the original structure, but 1-D CNNs in encoding-decoding layers (which are necessary for capturing the wide-range structures) causes this degradation, so we use 1-D convolution in the feature transformation block (i.e., DRN-SA blocks). 1×1 convolution is used to adjust the channel dimension before or after reshaping the feature map. In a forward pass, the source speech features are first projected and compressed into higher-level representation using two convolutional encoding layers. Either encoding layer here takes 2-D convolution, followed by instance normalization, and a GLU. The channel sizes per layer increase so that the feature depths get larger as the widths get narrower. In each 2-D CNN, we use the same padding to produce the output with the same resolution as the input. Instance normalization is used in each layer to improve the performance and stability of generators.

Subsequently, the compressed features are fed into the transformation block which consists of six DRN-SA blocks, in which 1-D casual dilated CNN is employed, followed by instance normalization (IN), GLUs, residual connection, and self-attention layers. This architecture enables the DRN-SA blocks can exploit the long-term relative dependencies of elements with different positions in compressed feature maps, progressively capturing the contextual relationship in enhancement procedure. In each self-attention layer of DRN-SA blocks, all of the keys, values and queries come from the same sequence, which is the output of the previous dilated residual layers. Each position in the self-attention layers can attend to all positions in the same compressed feature map.

The dilated residual network (DRN) is developed from [33] by using dilated operation in the residual network [37]. With dilation, the network can compensate for the reduction of the receptive field, proving a strong ability to incorporate larger local context. Before our transformation block, we reshape the compressed feature maps to 1-D, thus more effectively capturing the overall relationship. Our transformation block consists of six successive DRN-SA blocks. In the proposed DRN-SA block, for i^{th} layer in the residual block B_i^j $(j = 1, 2, 3, 4, 5, 6)$, the output of each dilated convolution layer is calculated by,

$$\left(B_i^j * u_i^j[k]\right)(i) = \sum\nolimits_{k=1}^{K} x\,(i + r \cdot k) u_i^j[k] \tag{15}$$

where r is the dilation rate and $u_i^j[k]$ is a filter of kernel size K. We use dilation rates $r = 1, 1, 2, 2, 4, 4$ in six successive residual blocks, respectively.

Figure 5 illustrates the dilation operation and conventional convolution operation on 1-D input features. In the conventional CNN structure, the receptive field size is thirteen after six successive convolutional layers (kernel size $= 3$, stride $= 1$). As shown in Figure 5, when using increasing

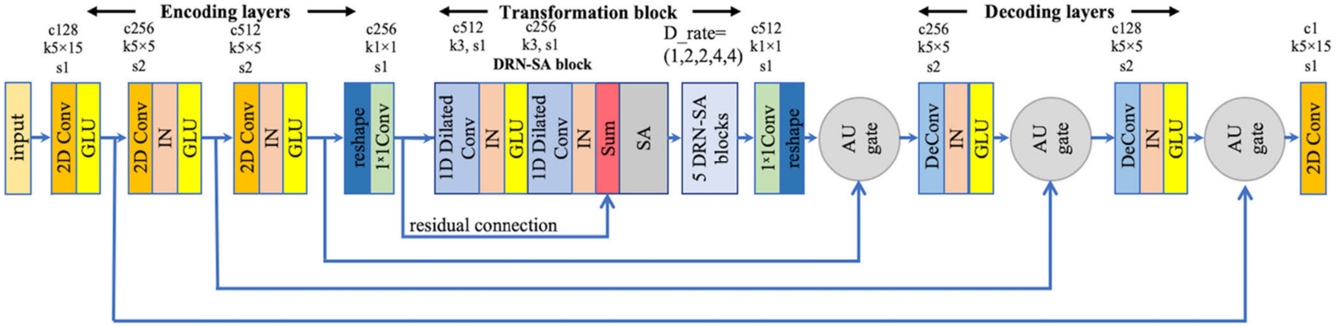


FIGURE 4. The network architecture of generators. In convolutional layers, c, k, and s denote numbers of output channels, kernel size, and stride, respectively. IN, GLU, SA, and AU gate indicate instance normalization, gated linear unit, self-attention, and attention U-net gate, respectively. 5 DRN-SA blocks denote five successive dilated residual self-attention blocks with increasing dilation rate $r = 1, 2, 2, 4, 4$.

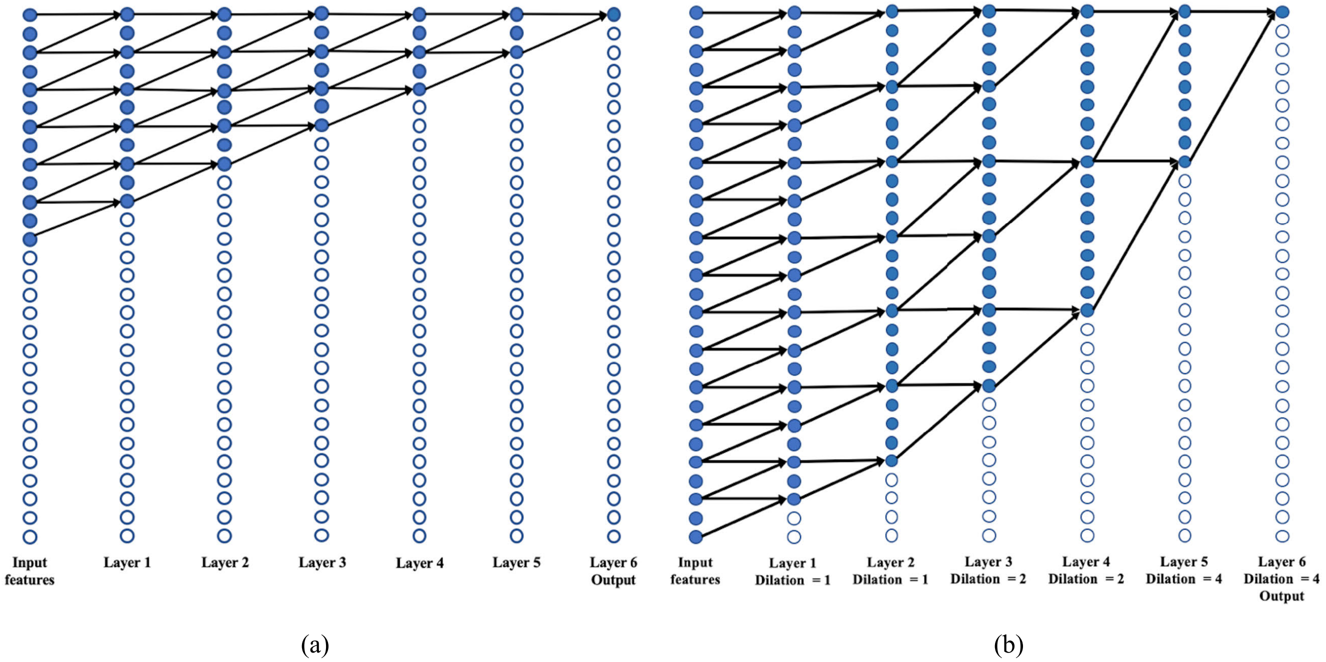


FIGURE 5. (a) Six layers CNN using 1D conventional convolution operation. (b) Six layers dilated CNN with increasing dilation rates ($r = 1, 1, 2, 2, 4, 4$).

TABLE 1. The detailed parameters of the six DRN-SA blocks, as well as the receptive field of each layer.

DRN-SA block	Block 1		Block 2		Block 3		Block 4		Block 5		Block 6	
Kernel size	3	3	3	3	3	3	3	3	3	3	3	3
Filters	512	256	512	256	512	256	512	256	512	256	512	256
Dilation rate	1	1	1	1	2	2	2	2	4	4	4	4
Receptive field	3×3	5×5	7×7	9×9	13×13	17×17	21×21	25×25	33×33	41×41	49×49	57×57

dilated rates in six dilated CNNs, the receptive field size increases to 29. In our DRN-SA blocks, the causal dilated convolutional layers are used twice per residual block, thus enlarging the size of the receptive field to 57 in the last convolutional layer. The parameters of six DRN-SA blocks are shown in Table 1.

Finally, the decoding layers reverse the encoding stage by transposed CNNs and reconstruct the target speech features. Here we use AU gates in decoding layers so that the encoding layers directly pass the salient information of source speech features to the decoding stage by attention gates. In the AU gates, the queries Q come from the output of the previous

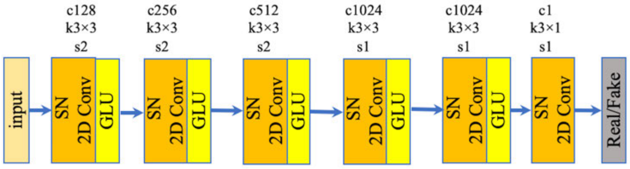


FIGURE 6. The network architecture of discriminators. SN and GLU are short for spectral normalization and the gated linear unit, respectively.

decoding layer or DRN-SA blocks, and the memory keys K and values V come from the output of the homologous encoding layers. This allows every position in the decoder to attend overall positions in the input features. In this way, the decoding layers to preserve source linguistic information in the reconstruction stage. Moreover, the U-net structure enables the gradients to flow deeper through the whole network [17], [31], and achieves more effective training.

C. SN PATCH DISCRIMINATOR

The proposed discriminator is illustrated in Figure 6. The generated features and target features are transformed into high-level representations, using 2-D CNNs with spectral normalization and GLUs. In previous GAN-based speech processing models [16], [22], the conventional discriminator has been extensively used, in which the last layer is fully connected. However, it requires more parameters to achieve wide-range receptive fields, thus, it is computationally expensive in training. Inspired by previous studies in image-to-image translation [38], [39], we designed our discriminator as Markovian (Patch) discriminator which penalizes the structure at the scale of a patch.

This patch discriminator replaces the last fully connected layer with a convolution layer and discriminates the realness of features on each $N \times N$ patch.

IV. EXPERIMENTS SETUP

A. DATASET

We evaluate our proposed method on the same dataset as proposed in [30]. We choose this public dataset because it has various types of non-stationary noise and we can compare our results with other published work. The dataset is a selection of the Voice Bank corpus [40] with 30 speakers. The training dataset includes 28 speakers' 11572 utterances in the same accent region (England), while the test set contains 2 speakers' (one male and one female) 824 utterances. The total duration of the training set is around 10 hours and the duration of the test set is around 30 mins. For both the training and testing sets, the average speech signal length was three seconds.

For the training set, audio samples are added with one of the 10 noise types (2 artificial and 8 from the DEMAND database [41]) at four signal-to-noise ratios (SNRs) of 0, 5, 10 and 15 dB. The noise is from different environments including offices, public spaces, transportation stations, and streets. The test set is created with 5 test-noise types (all from the DEMAND database, but totally unseen in the training set) at SNRs of 2.5, 7.5, 12.5 and 17.5 dB. The five types of chosen noise are living room, office, bus, and street noise. In our experiment, the utterances are down-sampled from 48 kHz to 16 kHz, so that the dataset size fits better for the speech enhancement task.

B. FEATURE PREPROCESSING

The WORLD speech synthesis system [42] is used to extract Mel-cepstral coefficients (MCEPs), logarithmic fundamental frequency (log F0), and aperiodicities (APs) at a frequency of 5 ms. The MCEPs and log F0 features are concatenated and form a 36-dimensional input acoustic feature vector to our model. The acoustic parameters are much easier to enhance using our models than the raw audio. After the enhancing process, MCEPs and log F0 are separated from the concatenated output acoustic features to reconstruct the enhanced speech waveform with APs by the WORLD vocoder. Aperiodicities are directly used without any enhancement because modifying APs does not significantly affect speech quality [43]. To randomize each batch, we randomly crop a fixed-length segment (128 frames) from a randomly selected audio file as the input sequence.

C. IMPLEMENTATION DETAILS

The network architectures are shown in Figure 4 and Figure 6. The detailed parameters of generators and discriminators are shown in Table 2. The number of filters, kernel size and stride of each convolutional layer in encoding-decoding layers of generators are: [128, (5, 15), 1], [256, (5, 5), 2], [512, (5, 5), 2], [256, (5, 5), 2], [128, (5, 5), 2], and [1, (5, 15), 1]. The patch discriminators use six 2-D convolutional layers with the number of filters, kernel sizes and strides as follows: [128, (3, 3), 1], [256, (3, 3), 2], [512, (3, 3), 2], [1024, (6, 6), 1], [1024, (3, 3), 1], and [1, (1, 3), 1].

We adopt the Adam optimizer [44] with the momentum term $\beta 1 = 0.5$ and train the networks with an initial learning rate of 0.0002 for discriminators and 0.0001 for generators, respectively. The same learning rates are maintained for the first 35 epochs, while they linearly decay in the remaining iterations. We set the batch size to 1 and use $\mathcal{L}_{identity}$ only for the first 2×10^5 iterations to guide the enhancement process. The λ_{cycle} and λ_{id} are set to 5 and 10 for the best performance, respectively. TensorFlow 1.14.0 is employed to implement the proposed framework as well as the baseline systems.

V. OBJECTIVE EVALUATION

A. EVALUATION METRICS

We use the following objective metrics to evaluate speech enhancement performance [45]. The metrics measure the similarity between the enhanced signal and the clean reference of the test set files. Higher values of all metrics indicate better speech performance.

1) PESQ

Perceptual evaluation of speech quality (PESQ) score [46] is the most common metric to evaluate speech quality,

TABLE 2. The parameter setup of the proposed networks.

Components	Input size (T×N×channel)	Kernel size	Stride	Filters	Output shape
2-1-2D Generators G and F					
Encoding layers	(128, 36, 1)	(5, 15)	1	128	(128, 36, 128)
	(128, 36, 128)	(5, 5)	2	256	(64,18,256)
	(64, 18, 256)	(5, 5)	2	512	(32, 9, 512)
Reshape	(32, 9, 512)	--	--	--	(288, 512)
1×1 Conv	(288, 512)	(1, 1)	1	256	(288, 256)
Transformation block (6 DRN-SA blocks)	$\begin{Bmatrix}(288, 256)\\(288, 512)\end{Bmatrix} \times 6$	$\begin{Bmatrix}3\\3\end{Bmatrix} \times 6$	1	$\begin{Bmatrix}512\\256\end{Bmatrix} \times 6$	$\begin{Bmatrix}(288, 512)\\(288, 256)\end{Bmatrix} \times 6$
1×1 Conv	(288, 256)	(1, 1)	1	512	(288, 512)
Reshape	(288, 512)	--	--	--	(32, 9, 512)
Decoding layers	(32, 9, 512)	(5, 5)	2	256	(64, 18, 256)
	(64, 18, 256)	(5, 5)	2	128	(128, 36, 128)
	(128, 36, 128)	(5, 15)	1	1	(128, 36, 1)
SN Patch Discriminators					
Down-sampling	(128, 36,1)	(3, 3)	2	128	(64, 18, 128)
	(64, 18, 128)	(3, 3)	2	256	(32, 9, 256)
	(32, 9, 256)	(3, 3)	2	512	(16, 5, 512)
Layer 4	(16, 5, 512)	(3, 3)	1	1024	(16, 5, 1024)
Layer 5	(16, 5, 1024)	(3, 3)	1	1024	(16, 5, 1024)
Output layer	(16, 5, 1024)	(3, 1)	1	1	(16, 5, 1)

especially using the wide-band version recommended in ITU-T P.862.2. *PSEQ* is a weighted sum of the average disturbance d_{sym} and the average asymmetrical disturbance d_{sym}, which can be defined as follows,

$$PSEQ = \alpha_0 + \alpha_1 \cdot d_{sym} + \alpha_2 \cdot d_{sym} \tag{16}$$

where $\alpha_0 = 4.5$, $\alpha_1 = -0.1$ and $\alpha_2 = -0.0309$. The PESQ value ranges from -0.5 to 4.5.

2) STOI

Short-Time Objective Intelligibility (STOI) [47] is used as a robust measurement index for nonlinear processing of noisy speech, e.g., noise reduction on speech intelligibility. The value of STOI is between zero and one, in which the score closer to one indicates higher intelligibility.

3) CSIG

The mean opinion score (MOS) prediction of the speech signal distortion, using a five-point scale [45].

4) CBAK

The MOS prediction of the intrusiveness of background noise, ranging from 1 to 5 [45].

5) COVL

The MOS prediction of the overall effect ($1 =$ bad, $2 =$ poor, $3 =$ fair, $4 =$ good, $5 =$ excellent) [45].

B. COMPARISON TO STATE-OF-THE-ART APPROACHES

Our method is compared with a variety of representative baselines (see below). All these baselines have been trained and evaluated on the same dataset in previous work, so we directly take the reported scores from those papers.

1) Wiener method based on a priori SNR estimation (Wiener) [7]: Wiener-filtering method reduces noise with low computational load for frequency-domain speech enhancement, based on a priori SNR estimation.
2) Speech Enhancement Generative Adversarial Network (SEGAN) [14]: SEGAN is the first speech enhancement approach based on the adversarial framework and works end-to-end with the raw audio. It applies to skip connection to generators, connecting each encoding layer to its homologous decoding layer. A high-frequency preemphasis filter is applied to all input speeches.
3) Time-frequency masking-based method using GAN along with L2 loss (MMSE-GAN) [16]: MMSE-GAN is a time-frequency masking-based enhancement

TABLE 3. The evaluation results of the proposed model compared with eight state-of-the-art approaches. Some metric scores are missing in this table, because they are not reported in the original papers. The best results obtained are highlighted in bold font.

Method	Features type	PESQ	STOI	CSIG	CBAK	COVL
Unprocessed	--	1.97	0.91	3.35	2.44	2.63
Wiener [7]	spectrogram	2.22	--	3.23	2.68	2.67
SEGAN [14]	waveform	2.16	0.93	3.48	2.94	2.80
Wave-U-Net [48]	waveform	2.40	--	3.54	3.24	2.96
MMSE-GAN [16]	spectrogram	2.53	0.93	3.80	3.12	3.14
WGAN-GP [15]	waveform	2.54	0.93	--	2.93	3.02
SERGAN [15]	waveform	2.62	**0.94**	--	--	--
DFL-SE [49]	waveform	--	--	3.86	**3.33**	3.22
MetricGAN [50]	spectrogram	**2.86**	--	3.99	3.18	3.42
Proposed	MCEPs+F0	2.72	**0.94**	**4.20**	3.21	**3.47**

approach based on GAN and learns the mask implicitly while predicting the clean T-F representation. It introduces a regularized objective function with the use of Mean Square Error (MSE) between predicted and target spectrum to overcome the failure of vanilla GAN in predicting the accurate mask.

4) Improved Speech Enhancement with the Wave-U-Net (Wave-U-Net) [48]: This approach uses the Wave-U-Net architecture for speech enhancement. It performs end-to-end audio source separation directly in the time domain.
5) GAN-based speech enhancement with Wasserstein loss and Gradient penalty (WGAN-GP) [15]: WGAN-GP employs Wasserstein losses and gradient penalty in the GANs' framework for speech enhancement. This method also operates at the time-domain waveform level and has a similar structure to SEGAN.
6) Speech Enhancement using Relativistic Generative Adversarial Network (SERGAN) [15]: SERGAN introduces relativistic GANs with a relativistic cost function at its discriminators and uses gradient penalty to improve speech enhancement performance in time-domain.
7) Speech Denoising with Deep Feature Losses (DFL-SE) [49]: DFL-SE proposes an end-to-end speech denoising approach and processes at the raw waveform level directly. This approach trains a fully-convolutional context aggregation network using a deep feature loss, which is based on comparing the internal feature activations in a different network.
8) Generative Adversarial Networks based Black-box Metric Scores Optimization (MetricGAN) [50]: Recently, some speech enhancement algorithms based on multi-objective loss functions have demonstrated great performance [50]–[52]. MetricGAN introduces an aim to optimize the generator with respect to one or multiple evaluation metrics such as PESQ and STOI, thus guiding the generators in GANs to generate data with improved metric scores. PESQ is used as the optimized metric in the loss function of discriminators.

Table 3 presents these metric values for our approach and the baselines, evaluated over the test set (824 utterances). By employing the relativistic loss, multi-attention, and dilated residual networks, our method outperforms all the baselines in terms of CSIG, COVL, and STOI measures, proving that our method focuses on reducing speech distortion, as well as improving speech intelligibility and overall signal quality.

As illustrated in Table 3, our method reduces the speech distortion (CSIG) by 5.3% and the background noise intrusiveness (CBAK) by 2.2%, while improving the overall signal quality (COVL) by 1.8% with respect to the MetricGAN method, which is the best previous GAN-based method. In terms of PESQ, our model does not perform as well as MetricGAN. This is because MetricGAN is based on a function approximation of objective sound quality assessments (OSQA) such as PESQ scores, in which PESQ score is approximated by using the discriminator D. Thus, the generators can generate speech samples with improved PESQ values than our model. However, as for other metrics, our method significantly outperforms MetricGAN model. Compared with the deep feature losses model (DFL-SE), the CSIG and COVL values of our enhanced speech are respectively increased by 8.8% and 8.1%, whilst providing near CBAK scores, proving our method focuses on reducing speech distortion instead of background noise.

C. ABLATION STUDY

To further discuss the contribution of each proposed component (relativistic losses, multi-attention, dilation operation, and spectral normalization), ten comparison systems based on CycleGAN are implemented (see below). All the extensions employ spectral normalization in discriminators and use dilation operation in residual blocks to further improve speech enhancement performance. In this experiment, the log-likelihood ratio (LLR) is used to measure the

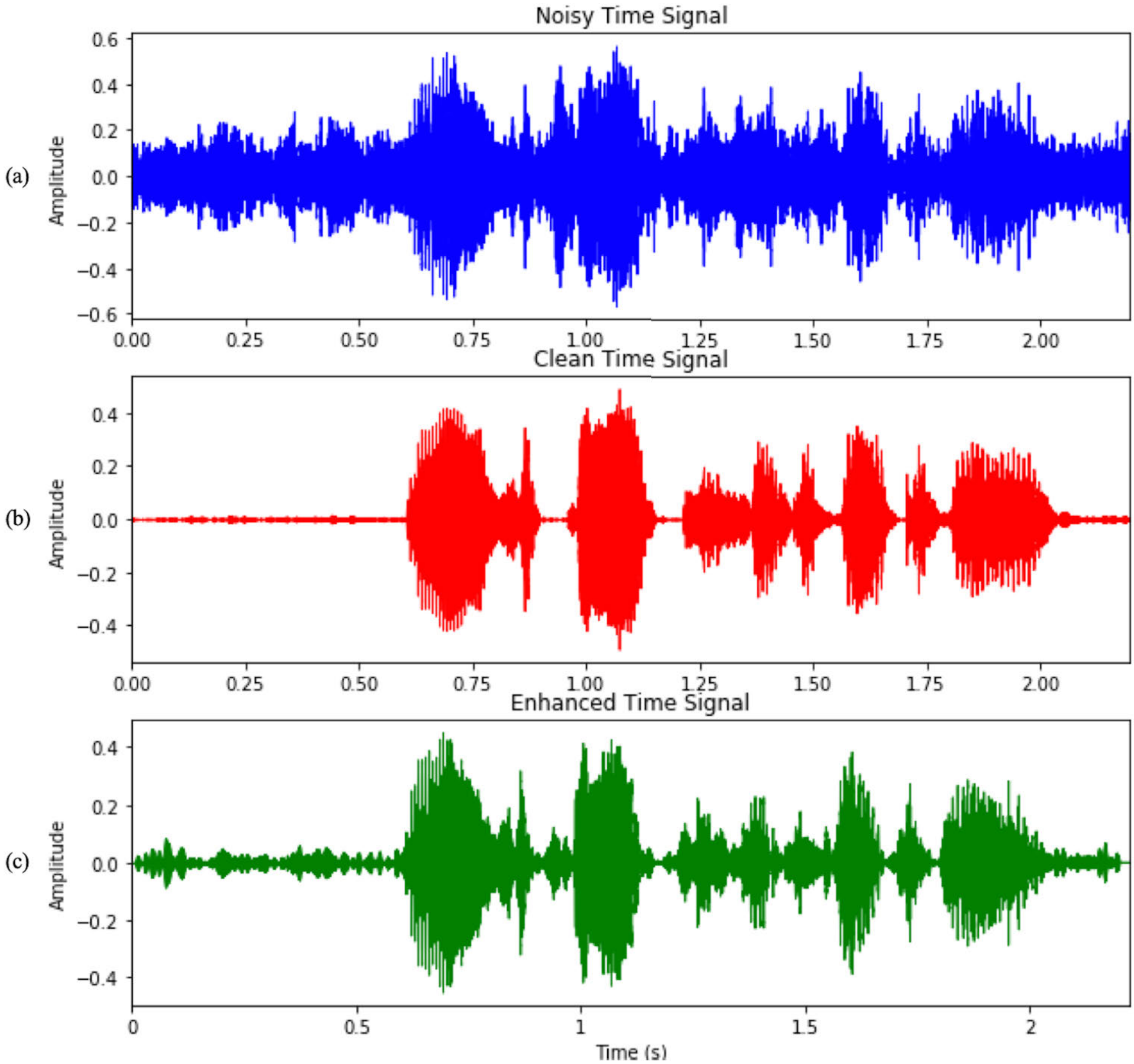


FIGURE 7. An example of speech enhancement results using our method. (a), (b) The blue lines represent the noisy speech time signal, while the red lines represent the clean speech time signal. (c) The corresponding enhanced speech time signal.

similarity between the enhanced signal and the clean signal, of which lower values indicate better performance.

1) Cycle-consistent GAN for speech enhancement (CycleGAN): CycleGAN is the baseline system for our task. In this model, the generator has a similar structure to Figure 3, including encoding-decoding layers and the transformation block. However, it does not employ an attention mechanism or U-net structure and only uses residual blocks as the transformation block.
2) Relativistic CylcleGAN for speech enhancement (CRGAN): Relativistic least-square average loss is employed to improve the quality of the generated features.
3) CRGAN with self-attention layers (SA-CRGAN): Self-attention layers are employed in residual block to model relative dependencies in feature transformation.
4) CRGAN with attention U-net gates (AU-CRGAN): Attention gates are employed in U-net encoding-decoding layers to preserve source linguistic information in the reconstruction stage and alleviate the loss of speech structure.
5) CRGAN with a multi-attention mechanism (MA-CRGAN): Multi-attention mechanism composed of the combination of attention u-net gates and self-attention is employed to further enhance the system performance.

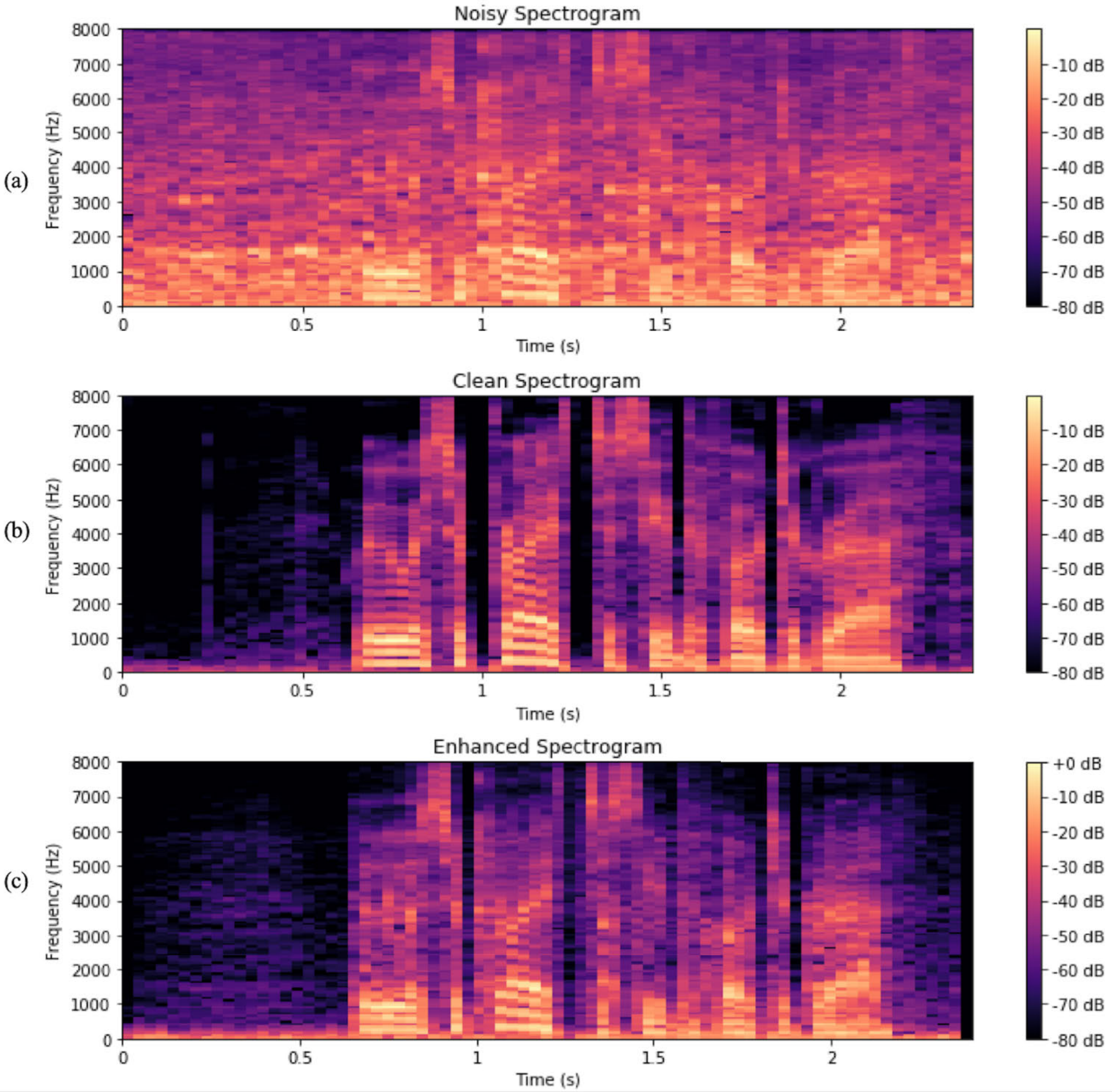


FIGURE 8. The enhancement results for the utterance of p232_052.wav in the test set with SNR = 2.5 dB. The spectrogram of (a) the noisy speech (b) the clean reference speech (c) The enhanced speech by our proposed method.

6) Improved CycleGAN (iCycleGAN), improved CRGAN (iCRGAN), improved CRGAN-SA (iSA-CRGAN), improved AU-CRGAN (iAU-CRGAN), and the proposed iMA-CRGAN: Spectral normalization and dilated residual networks are employed in each baseline system. Spectral normalization is used to stabilize GAN training in discriminators and dilation operation is used to increase the receptive field of each residual convolution layer in the transformation block.

As shown in Table 4, by employing spectral normalization in discriminators and dilation operation in transformation block, all improved models have gained better performance on PESQ, CSIG, CBAK, and COVL, while achieving similar STOI and LLR scores. By applying relativistic average least-square losses, CRGAN has gained +2.5%, +1.6%, and +9.6% relative improvement respectively on PESQ, CSIG, and COVL over the baseline CycleGAN system with least-square loss functions, suggesting that relativistic discriminators lead to better generated samples. Self-attention in the transformation block results in better performance, especially in reducing speech distortion (CSIG) and improving overall signal quality (COVL). Incorporating AU gates

TABLE 4. The evaluation results of the ten comparison models. All the extension models are all implemented by ourselves. IN and SN represent instance normalization and spectral normalization in discriminators, respectively.

Method	Normalization	Dilated	PESQ	STOI	CSIG	CBAK	COVL	LLR
Unprocessed	--	--	1.97	0.91	3.35	2.44	2.63	0.46
CycleGAN	IN	No	2.32	0.92	3.75	2.93	3.02	0.44
iCycleGAN	SN	Yes	2.36	0.92	3.78	2.96	3.08	0.43
CRGAN	IN	No	2.38	0.93	3.81	3.02	3.04	0.41
iCRGAN	SN	Yes	2.41	0.93	3.86	3.04	3.12	0.39
SA-CRGAN	IN	No	2.49	0.93	4.02	3.10	3.23	0.38
iSA-CRGAN	SN	Yes	2.53	0.93	4.07	3.11	3.26	0.38
AU-CRGAN	IN	No	2.56	0.93	4.06	3.14	3.33	0.36
iAU-CRGAN	SN	Yes	2.61	0.94	4.12	3.16	3.39	0.35
MA-CRGAN	IN	No	2.69	0.94	4.15	3.19	3.45	0.33
Proposed	SN	Yes	**2.72**	**0.94**	**4.20**	**3.21**	**3.47**	**0.32**

and self-attention in generators as multi-attention mechanism yields further improvements over the SA-CRGAN model and the AU-CRGAN model. With the combining usage of relativistic losses, AU gates, DRN-SA blocks, and spectral normalization, the proposed model achieves the best performance, consistently improving all metrics by a significant margin compared with other implemented models.

Figure 7 and Figure 8 illustrate the speech enhancement result of a test sample (p232_052.wav) with a duration of 2.199 seconds under a low signal-to-noise (SNR = 2.5 dB) condition. The noisy speech signal comes from a male speaker in background non-stationary noise (“cocktail party” and musical noise). As shown in Figure 7, the enhanced time signal filters out the background noise without the speaker’s voice, while tracking the clean speech when the foreground speaker starts talking. Figure 8 also demonstrates the enhanced spectrogram significantly reduces the background noise in the first half of utterance dominated by the noise components and preserves the texture and outline of the speech in the second half, in which the noise components and the speech components are not distinguished. In summary, the proposed method filters out the noise components and retain the speech components, thus better restoring the enhanced speech signal.

VI. CONCLUSION AND FUTURE WORK

Considering the impressive performance of Cycle-GANs in domain transformation, this article investigates an improved cycle-consistent relativistic GAN as feature mapping networks for speech enhancement. To stabilize GAN training, we employ the relativistic average least-square GAN (RaLSGAN) loss and spectral normalization. Then, we propose a multi-attention mechanism in gated-convolutional U-shape generators to preserve linguistic information and exploit long-term dependency, whilst using dilation operation in residual blocks to enlarge the receptive field size in feature transformation. Our proposed model achieves state-of-the-art performance and significantly outperforms WIENER filtering, SEGAN, Wave-U-Net, MMSEGAN, SERGAN, DFL-SE, and MetricGAN on most objective metrics. Experiments also show that our method focuses on reducing speech distortion (CSIG) while improving overall speech effect (COVL) and speech intelligibility (STOI) in the enhancement process.

In the future, we will apply this architecture in complex spectral mapping for speech enhancement and research a more effective attention mechanism with less computational cost.

REFERENCES

[1] T. Ochiai, S. Watanabe, T. Hori, and J. Hershey, “Multichannel end-to-end speech recognition,” in *Proc. ICML*, Sydney, NSW, Australia, Aug. 2017, pp. 2632–2641.

[2] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, “An overview of noise-robust automatic speech recognition,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 4, pp. 745–777, Apr. 2014.

[3] G. S. Bhat, N. Shankar, C. K. A. Reddy, and I. M. S. Panahi, “A real-time convolutional neural network based speech enhancement for hearing impaired listeners using smartphone,” *IEEE Access*, vol. 7, pp. 78421–78433, 2019.

[4] W. Yuan and B. Xia, “A speech enhancement approach based on noise classification,” *Appl. Acoust.*, vol. 96, pp. 11–19, Sep. 2015.

[5] B. M. Mahmmod, A. R. Ramli, T. Baker, F. Al-Obeidat, S. H. Abdulhussain, and W. A. Jassim, “Speech enhancement algorithm based on super-Gaussian modeling and orthogonal polynomials,” *IEEE Access*, vol. 7, pp. 103485–103504, 2019.

[6] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[7] P. Scalart and J. V. Filho, “Speech enhancement based on *a priori* signal to noise estimation,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Conf.*, May 1996, pp. 629–632.

[8] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[9] Y. Wang, A. Narayanan, and D. Wang, “On training targets for supervised speech separation,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[10] A. Nicolson and K. K. Paliwal, “Deep learning for minimum mean-square error approaches to speech enhancement,” *Speech Commun.*, vol. 111, pp. 44–55, Aug. 2019.

[11] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. INTERSPEECH*, Lyon, France, Aug. 2013, pp. 436–440.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, Montpellier, QC, Canada, 2014, pp. 2672–2680.

[13] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 1, pp. 84–96, Jan. 2018.

[14] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," in *Proc. Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 3642–3646.

[15] D. Baby and S. Verhulst, "Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 106–110.

[16] M. H. Soni, N. Shah, and H. A. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 5039–5043.

[17] X. Hao, X. Su, Z. Wang, H. Zhang, and A. Batushiren, "UNetGAN: A robust speech enhancement approach in time domain for extremely low signal-to-noise ratio condition," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 1786–1790.

[18] S. Pascual, J. Serrà, and A. Bonafonte, "Towards generalized speech enhancement with generative adversarial networks," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 1791–1795.

[19] J.-C. Chou, C.-C. Yeh, H.-Y. Lee, and L.-S. Lee, "Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations," in *Proc. Interspeech*, Graz, Austria, Sep. 2018, pp. 501–505.

[20] Z. Meng, J. Li, Y. Gong, and B.-H. Juang, "Adversarial teacher-student learning for unsupervised domain adaptation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 5949–5953.

[21] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2223–2232.

[22] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Roma, Italy, Sep. 2018, pp. 6820–6824.

[23] Z. Meng, J. Li, Y. Gong, and B.-H. Juang, "Cycle-consistent speech enhancement," in *Proc. Interspeech*, Hyderabad, India, Sep. 2018, pp. 1165–1169.

[24] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Cyclegan-VC2: Improved cyclegan-based non-parallel voice conversion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 6820–6824.

[25] A. Jolicoeur-Martineau, "The relativistic discriminator: A key element missing from standard GAN," 2018, *arXiv:1807.00734*. [Online]. Available: http://arxiv.org/abs/1807.00734

[26] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2794–2802.

[27] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," 2018, *arXiv:1802.05957*. [Online]. Available: http://arxiv.org/abs/1802.05957

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst., Annu. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 6000–6010.

[29] T. Lan, Y. Lyu, W. Ye, G. Hui, Z. Xu, and Q. Liu, "Combining multi-perspective attention mechanism with convolutional networks for monaural speech enhancement," *IEEE Access*, vol. 8, pp. 78979–78991, 2020.

[30] X. Gao, F. Feng, X. He, H. Huang, X. Guan, C. Feng, Z. Ming, and T.-S. Chua, "Hierarchical attention network for visually-aware food recommendation," *IEEE Trans. Multimedia*, vol. 22, no. 6, pp. 1647–1659, Jun. 2020.

[31] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Oct. 2015, pp. 234–241.

[32] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," 2018, *arXiv:1805.08318*. [Online]. Available: http://arxiv.org/abs/1805.08318

[33] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 636–644.

[34] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech," in *Proc. 9th ISCA Speech Synth. Workshop*, Sunnyvale, CA, USA, Sep. 2016, pp. 146–152.

[35] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. ICML*, Sydney, NSW, Australia, Aug. 2017, pp. 933–941.

[36] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*. [Online]. Available: http://arxiv.org/abs/1607.08022

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[38] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1874–1883.

[39] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jul. 2017, pp. 5967–5976.

[40] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Proc. Int. Conf. Oriental COCOSDA Held Jointly Conf. Asian Spoken Lang. Res. Eval. (O-COCOSDA/CASLRE)*, Gurugram, India, Nov. 2013, pp. 1–4.

[41] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multichannel acoustic noise database: A database of multichannel environmental noise recordings," *J. Acoust. Soc. Amer.*, vol. 133, no. 5, p. 3591, 2013.

[42] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. E99.D, no. 7, pp. 1877–1884, 2016.

[43] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," in *Proc. INTERSPEECH*, New York, NY, USA, Sep. 2006, pp. 2266–2269.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, 2015, pp. 1–15.

[45] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang., Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[46] 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 *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Salt Lake City, UT, USA, May 2001, pp. 749–752.

[47] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Dallas, TX, USA, 2010, pp. 4214–4217.

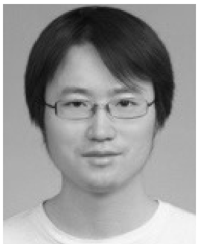
[48] C. Macartney and T. Weyde, "Improved speech enhancement with the Wave-U-Net," 2018, *arXiv:1811.11307*. [Online]. Available: http://arxiv.org/abs/1811.11307

[49] F. G. Germain, Q. Chen, and V. Koltun, "Speech denoising with deep feature losses," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 2723–2727.

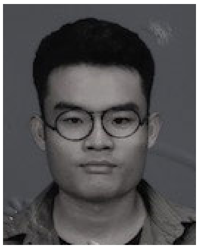
[50] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Proc. ICML*, Long Beach, CA, USA, Jun. 2019, pp. 2031–2041.

[51] M. Kolbæk, Z. Tan, S. H. Jensen, and J. Jensen, "On loss functions for supervised monaural time-domain speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang., Process.*, vol. 28, pp. 825–838, 2020.

[52] A. Azarang and N. Kehtarnavaz, "A review of multi-objective deep learning speech denoising methods," *Speech Commun.*, vol. 122, pp. 1–10, Sep. 2020.



YUTIAN WANG (Member, IEEE) received the B.S. and M.S. degrees in communication and information system from the Communication University of China, Beijing, in 2007, and the Ph.D. degree from the Key Laboratory of Media Audio and Video, Ministry of Education, Communication University of China, in 2012. From 2012 to 2014, he was a Research Assistant with the Key Laboratory of Media Audio and Video, Ministry of Education, Communication University of China. Since 2015, he has been an Assistant Professor with the Key Laboratory of Media Audio and Video, Ministry of Education. His research interests include signal processing theory and deep learning toward audio application, including speech enhancement, source separation, and speech synthesis.



GUOCHEN YU received the B.S. degree in communication engineering from the Communication University of China, Beijing, in 2017, where he is currently pursuing the Ph.D. degree with the Key Laboratory of Media Audio and Video, Ministry of Education. His research interests include deep learning, cycle-consistent generative adversarial networks, attention mechanism, music generation, and speech enhancement.

JINGLING WANG received the Ph.D. degree from the Communication University of China, Beijing, China, in 2011. She is currently a Professor with the Key Laboratory of Media Audio and Video, Ministry of Education, Communication University of China. Her research interests include computer vision, statistical modeling, and telecommunications.

HUI WANG (Member, IEEE) received the Ph.D. degree from the Communication University of China, Beijing, China, in 2011. He is currently a Professor with the State Key Laboratory of Media Convergence and Communication, Communication University of China. His research interests include audio signal processing, media convergence, sound field reproduction and broadcasting, and television technology.

QIN ZHANG received the Ph.D. degree from The University of British Columbia, Vancouver, BC, Canada, in 1990. He is currently a Professor and the Director of the Key Laboratory of Media Audio and Video, Ministry of Education, Communication University of China. He has served as a Senior Reviewer for Nature Science Foundation of China. His research interests include multimedia technology and next-generation coding scheme of image, video, and audio.