entropy

MDPI

Article

CGA-MGAN: Metric GAN Based on Convolution-Augmented Gated Attention for Speech Enhancement

Haozhe Chen [1,2,3] **and Xiaojuan Zhang** [1,2,*]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China
[2] Key Laboratory of Electromagnetic Radiation and Sensing Technology, Chinese Academy of Sciences, Beijing 100190, China
[3] School of Electronic, Electrical, and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: xjzhang@mail.ie.ac.cn

Abstract: In recent years, neural networks based on attention mechanisms have seen increasingly use in speech recognition, separation, and enhancement, as well as other fields. In particular, the convolution-augmented transformer has performed well, as it can combine the advantages of convolution and self-attention. Recently, the gated attention unit (GAU) was proposed. Compared with traditional multi-head self-attention, approaches with GAU are effective and computationally efficient. In this CGA-MGAN: MetricGAN based on Convolution-augmented Gated Attention for Speech Enhancement, we propose a network for speech enhancement called CGA-MGAN, a kind of MetricGAN based on convolution-augmented gated attention. CGA-MGAN captures local and global correlations in speech signals at the same time by fusing convolution and gated attention units. Experiments on Voice Bank + DEMAND show that our proposed CGA-MGAN model achieves excellent performance (3.47 PESQ, 0.96 STOI, and 11.09 dB SSNR) with a relatively small model size (1.14 M).

Keywords: CGA-MGAN; gated attention unit; speech enhancement

check for updates

Citation: Chen, H.; Zhang, X. CGA-MGAN: Metric GAN Based on Convolution-Augmented Gated Attention for Speech Enhancement. *Entropy* **2023**, *25*, 628. https://doi.org/10.3390/e25040628

Academic Editors: Deniz Gençağa and Boris Ryabko

Received: 9 January 2023
Revised: 15 March 2023
Accepted: 4 April 2023
Published: 6 April 2023

1. Introduction

Speech enhancement (SE) systems are usually used at the frontend of automatic speech recognition processes [1], communication systems [2], and hearing aids [3] to remove noise from speech. Methods based on traditional signal processing, such as subtraction [4], Wiener filtering [5], and minimum mean square estimation [6], are widely used in speech enhancement. Although these methods perform well in handling stationary noises, it is challenging to deal with nonstationary noises. With the development of deep neural networks (DNNs), this model has been used more frequently in speech enhancement in recent years.

The traditional time–frequency domain DNN model reconstructs the speech magnitude spectrum by estimating the mask function [7–9] or directly predicting the magnitude spectrum of clean speech [10], ignoring the role of phase information. However, phase information improves speech perception quality under a low signal-to-noise ratio (SNR) [11,12]. In [13], researchers proposed recovering magnitude and phase information simultaneously in the time–frequency (TF) domain by estimating the complex ratio mask (CRM) [14]. However, due to the compensation effect between the magnitude and phase [15], the simultaneous enhancement of magnitude and phase reduces the effect of magnitude estimation [16]. In [17], researchers proposed a decoupling-style phase-aware method. By building a two-path network, the magnitude is estimated to be increased first. Then, the spectrum is refined using residual learning, which can effectively alleviate the problem of the compensation effect.

The self-attention mechanism [18,19] can model the global context, but it is not good at extracting fine-grained local feature patterns. Convolution [20–22] is good at capturing local feature information but needs to improve its capture of global information. The convolution-augmented transformer (conformer) [23–25] combines convolution and self-attention and models local speech information by inserting deep convolution into the transformer, which can better extract the features of speech signals. However, the computational complexity of the network increases due to the large number of multi-head self-attention (MHSA) structures and feed-forward modules used in multiple stacked conformer blocks.

In speech enhancement, people usually care most about the quality or clarity of speech. Therefore, objective evaluation metrics considering human perception as the cost function can improve speech quality more directly. Nevertheless, standard metrics, such as the perceptual evaluation of speech quality (PESQ), are nondifferentiable. Therefore, they cannot be used directly as cost functions. MetricGAN can mimic evaluation functions such as PESQ by building neural networks. Considering that the objective function based on point distance may not fully reflect the perception difference between noisy and clean speech signals, Ref. [26] introduces MetricGAN [27] to the conformer, which allows the net to use the evaluation metric score learned by the metric discriminator to optimize the generator.

In this paper, we propose a convolution-augmented gated attention MetricGAN called CGA-MGAN, composed of a generator and a discriminator. We use the convolution-augmented gated attention unit (CGAU) to extract speech features in the generator. Compared with the conformer, which used MHSA, the CGAU model we propose allows the network to use weaker single-head self-attention (SHSA). CGAU can capture the local features and global information of speech simultaneously and minimize quality loss with faster speed, lower memory occupation, and better effect. The discriminator can estimate a black-box, nondifferentiable metric to guide the generator in enhancing speech.

Our main contributions can be summarized as the following points:

- We construct an encoder–decoder structure including gating blocks using the decoupling-style phase-aware method that can collaboratively estimate the magnitude and phase information of clean speech in parallel and avoid the compensation effect between magnitude and phase;
- We propose a convolution-augmented gated attention unit that can capture time and frequency dependence with lower computational complexity and achieve better results than the conformer;
- The proposed approach is superior to the previous approaches on the Voice Bank + DEMAND dataset [28], and an ablation experiment has verified our design choice.

The remainder of this paper is organized as follows: Section 2 introduces the related work of speech enhancement. Section 3 analyzes the specific architecture of the CGA-MGAN model we propose. Section 4 introduces the experimental setup, including the dataset used for the experiment, the network's training setup, and the experimental results' evaluation indicators. Section 5 analyzes the experimental results, compares them with some existing models, and conducts an ablation experiment. Finally, Section 6 summarizes this work and suggests some future research directions.

2. Related Works

This paper focuses on a convolution-augmented gated attention MetricGAN model for speech enhancement. In this section, we briefly introduce MetricGAN, conformers, and their basic working principles. In addition, we review the structure of the standard transformer and briefly introduce the basic principle of the new transformer variant, the gated attention unit (GAU).

2.1. MetricGAN

Before introducing MetricGAN, we will first introduce how to use the general GAN network for speech enhancement. GAN can simulate real data distribution by employing an alternative mini-max training scheme between the generator and the discriminator. By using the least-squares GAN method [29] to minimize the following loss function, we can train the generator to map noisy speech, x, to clean speech, y, and to generate enhanced speech.

$$L_{G(LSGAN)} = E_x\left[D(G(x), x) - 1)^2\right] \tag{1}$$

Here, G represents the generator and D represents the discriminator. By minimizing the following loss function, we can train D to distinguish between clean speech and enhanced speech:

$$L_{D(LSGAN)} = E_{x,y}\left[(D(y, x) - 1)^2 + (D(G(x), x) - 0)^2\right] \tag{2}$$

Here, E represents the expectation, 1 represents the clean speech, and 0 represents the enhanced speech. $G(x)$ represents the enhanced speech, and $D(\cdot)$ represents the output result of the discriminator.

MetricGAN consists of a generator and a discriminator, the same as the general GAN network. The generator enhances speech. The discriminator treats the objective evaluation function as a black box and trains the surrogate evaluation function. During training, the discriminator and the generator are updated alternately to guide the generator to generate higher-quality speech.

Unlike the general GAN network, MetricGAN introduces a function ($Q(I)$) to represent the normalized metric to be simulated. The loss functions of MetricGAN are shown in the following formulas:

$$L_{G(MetricGAN)} = E_x\left[(D(G(x), y) - s)^2\right] \tag{3}$$

$$L_{D(MetricGAN)} = E_{x,y}\left[(D(y, y) - Q(y, y))^2 + (D(G(x), y) - Q(G(x), y))^2\right] \tag{4}$$

Formula (3) is the loss function of the generator network, where s represents the expected distribution score. When s = 1, the generator will generate enhanced speech that is close to clean speech.

Formula (4) is the loss function of the discriminator network, where $Q(I)$ represents the function of the target evaluation metric normalized between 0 and 1, and I represents the speech pair to be evaluated. When the inputs are both clean voices, $I = (y, y)$; when the inputs are an enhanced voice and a corresponding clean voice, $I = (G(x), y)$.

The training process for MetricGAN can be condensed into the following schema:

- Input noisy speech, x, into the generator to generate enhanced speech, $G(x)$;
- Input a clean–clean speech pair, (y, y), into the discriminator to calculate the output, $D(y, y)$, and calculate $Q(y, y)$ through the objective evaluation function;
- Input an enhanced–clean speech pair, $(G(x), y)$, into the discriminator to calculate the output, $D(G(x), y)$, and calculate $Q(G(x), y)$ through the objective evaluation function;
- Calculate the loss function of the generator and the discriminator and update the weights of both networks.

2.2. Gated Attention Unit

Recently, Hua W. [30] proposed a new transformer variant called GAU. Compared with the standard transformer, it has faster speed, lower memory occupation, and a better effect.

The standard transformer comprises, alternately, an attention block and a feed-forward network (FFN) layer, which consists of two multi-layer perceptron (MLP) layers. The attention block uses MHSA, as shown in Figure 1a. Unlike the standard transformer, GAU has only one layer, which makes networks stacked with GAU modules simpler and easier to understand. GAU creatively uses the gated linear unit (GLU) instead of the FFN layer. The structure of the GLU is shown in Figure 1b. The powerful performance of GLU allows GAU to weaken its dependence on attention. GAU can use SHSA instead of MHSA, achieving the same or even better effects compared with the standard transformer [30]. It not only improves the computing speed but also reduces memory occupation. The structure of GAU is shown in Figure 2.

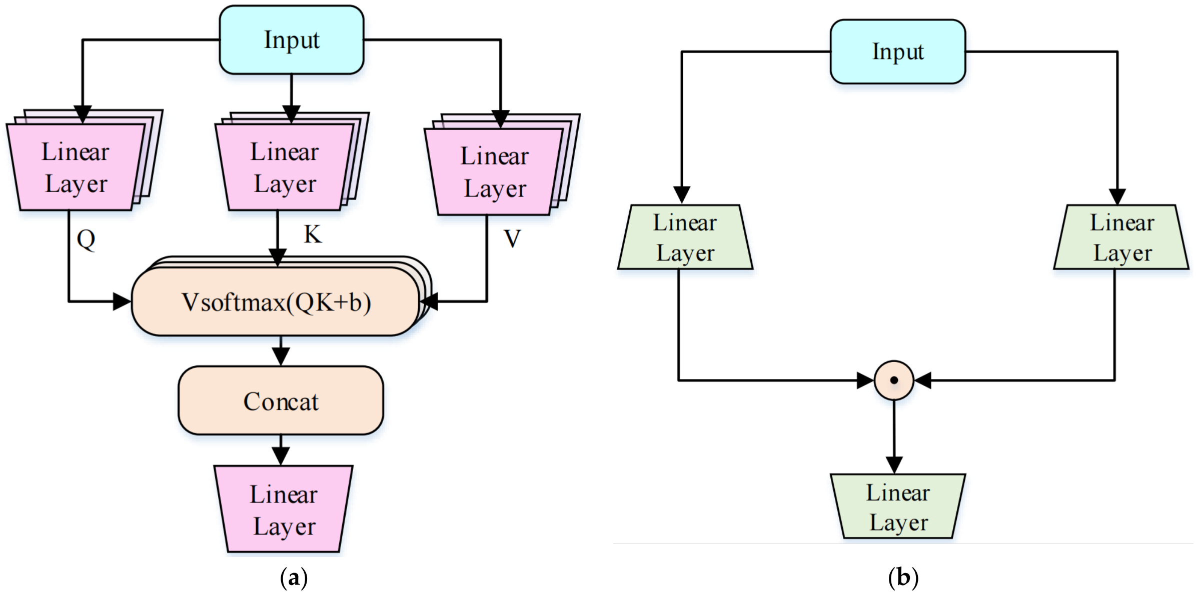

Figure 1. (**a**) Muti-head self-attention; (**b**) gated linear unit.

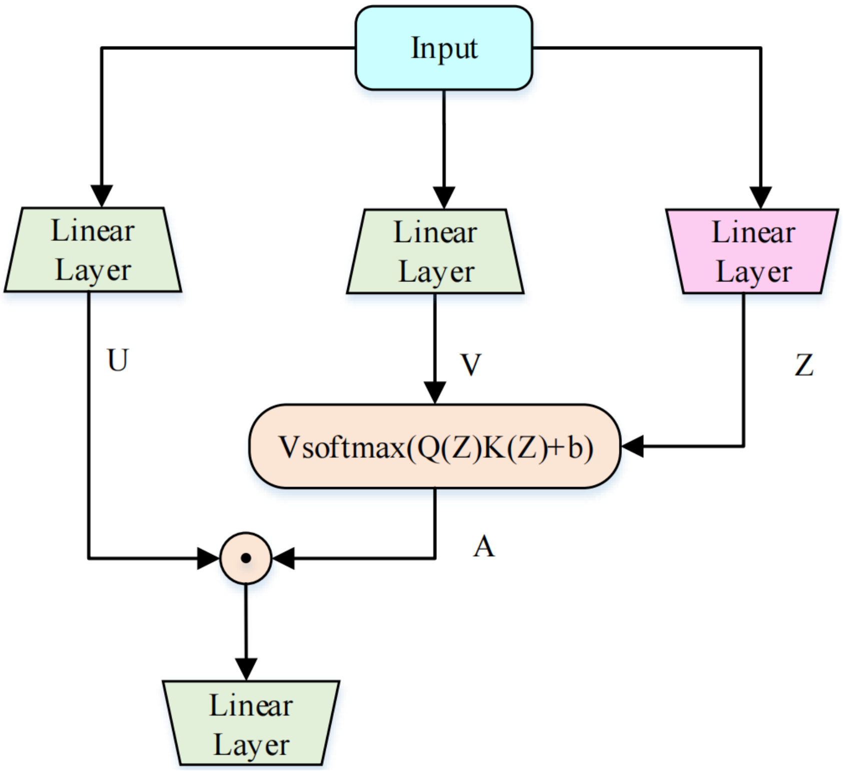

Figure 2. Gated attention unit.

2.3. Conformer

The conformer was first used in speech recognition and can also be used for speech enhancement. Since a pronunciation unit is composed of multiple adjacent speech frames, the convolution mechanism can better extract fine-grained local feature patterns, such as pronunciation unit boundaries. The conformer combines the convolution and self-attention modules and gives full play to their advantages. The main structure of the conformer is shown in Figure 3. The conformer block consists of four parts: the first feed-forward module, an MHSA module, a convolution module, and the second feed-forward module. The detailed structure of the convolution block is shown in Figure 4, and the detailed structure of the feed-forward module is shown in Figure 5. Inspired by Macaron Net [31], the conformer adopts a makaron-style structure. The convolution module and the MHSA module are placed between two feed-forward modules. By stacking the conformer blocks, speech features are extracted step by step to achieve speech recognition or speech enhancement.

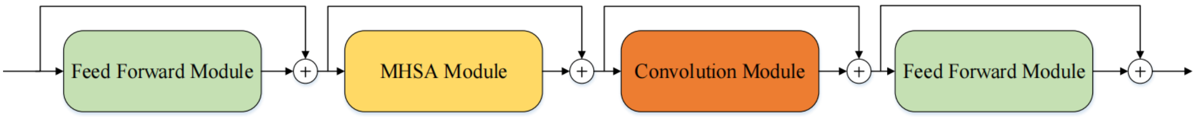

Figure 3. Conformer architecture.

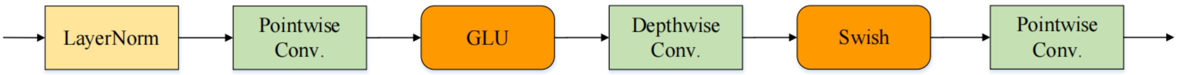

Figure 4. Detailed structure of the convolution block.

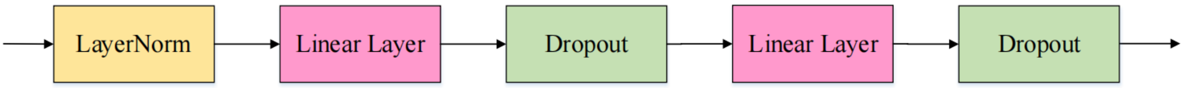

Figure 5. Detailed structure of the feed-forward module.

2.4. Limitations and Our Approach

MetricGAN is a great contribution to the application of GAN to speech enhancement. GAN simulates evaluation metrics that were originally nondifferentiable so that it can be used as a loss function. However, the performance of the MetricGAN generator limits its speech enhancement effect. In our CGA-MGAN model, we use the idea of MetricGAN to build a discriminator and a generator with an encoder–decoder structure, including gating blocks using the decoupling-style phase-aware method, which can greatly improve the network's speech enhancement performance.

In addition, although GAU has been applied to natural speech processing, previous research has yet to involve the field of speech enhancement. This paper is the first application of GAU to speech enhancement. Compared with makaron-style structures used in conformers, the CGAU we propose uses GLU to replace two feed-forward modules in the conformer, replaces MHSA with SHSA, and perfectly integrates the convolution module and GAU, significantly reducing the computational complexity of the network. A convolution-augmented GAU constructed this way can extract global and local features simultaneously to obtain better speech quality.

3. Methodology

In this section, we introduce the composition of the CGA-MGAN model, including the encoder–decoder structure of the generator, the structure of the two-stage CGA block, and the structure of the metric discriminator. Finally, we introduce the loss function of the generator and the metric discriminator.

The architecture of the generator is shown in Figure 6. The generator consists of an encoder–decoder structure, including gating blocks using the decoupling-style phase-aware method and four two-stage CGA blocks. First, it takes a discrete noisy signal, $y \in \mathbb{R}^{B \times N \times 1}$, with N samples as the input. Then, we convert the input signal to $Y_o \in \mathbb{R}^{B \times T \times F \times 1}$ in a time-frequency representation domain using short-time Fourier transform (STFT), where

T represents the number of frames and F represents the number of frequency bins of the complex spectrogram. After that, a power law compression with a compression exponent of 0.3 is applied to the spectrum:

$$Y = |Y_o|^c e^{jY_p} = Y_m e^{jY_p} = Y_r + jY_i \tag{5}$$

where c is the compression exponent. Then, the magnitude, Y_m, real component, Y_r, and imaginary component, Y_i, of the spectrum are concatenated as $Y_{in} = [Y_m; Y_r; Y_i] \in \mathbb{R}^{B\times T\times F\times 3}$ as the input of the encoder.

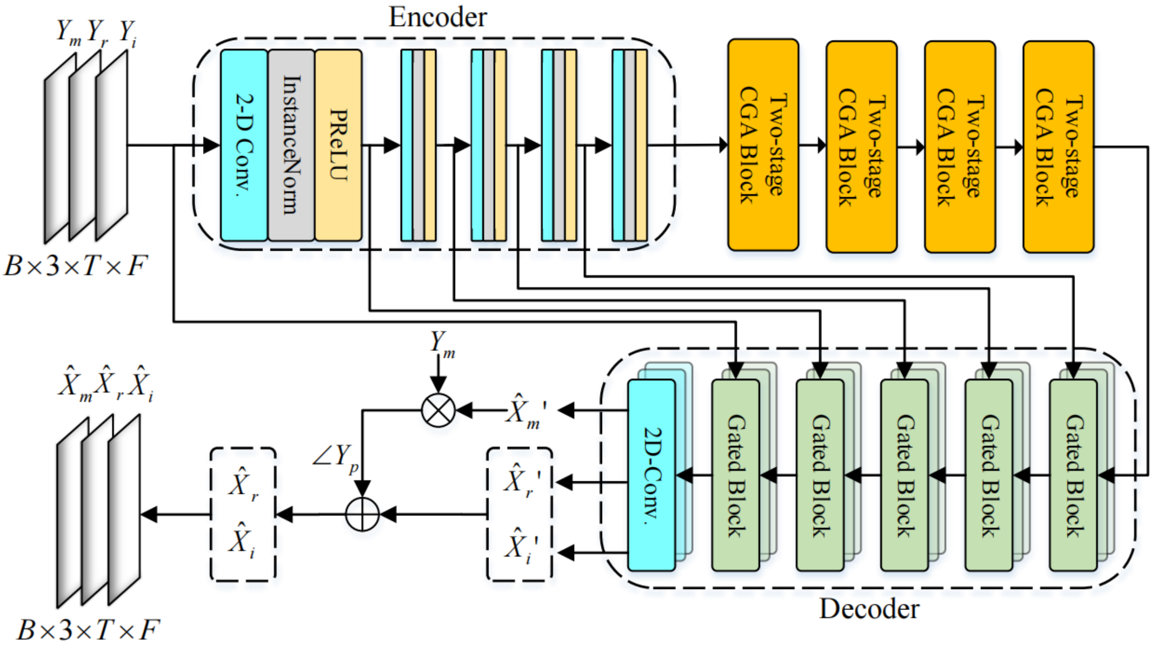

Figure 6. Encoder–decoder generator architecture.

3.1. Encoder and Decoder

The encoder of the generator consists of five encoder blocks with concatenation operations. The last encoder block halves the frequency dimension to reduce complexity. Each encoder block consists of a Conv2D (two-dimensional convolution) layer, an instance normalization layer, and a parameter ReLU (PReLU) activation layer. The output feature of the encoder is $Y_{enc} \in \mathbb{R}^{B\times T\times F'\times C}$, where $F' = F/2, C = 64$.

The decoder of the generator consists of three decoder blocks, including a magnitude mask estimation decoder block and two complex spectrum estimation decoder blocks, which output the multiplicative mask of the magnitude, $\hat{X}_m'$; the real component, $\hat{X}_r'$; and the imaginary component, $\hat{X}_i'$, of the spectrum in parallel. Each decoder block consists of five gated blocks and a Conv2D layer. The first gated block samples the frequency dimension up to F. The gated block consists of a Conv2D Transpose layer and two Conv2D blocks. The structure of the Conv2D block in the decoder block is the same as that of the encoder block. The gated block learns features from the encoder and suppresses its unwanted parts, which is shown in Figure 7. After five decoder blocks, the Conv2D layer compresses the number of channels to obtain $\hat{X}_m'$, $\hat{X}_r'$, $\hat{X}_i'$. Then, we multiply the $\hat{X}_m'$ and magnitude Y_m of noisy speech to obtain the preliminary estimated spectrum.

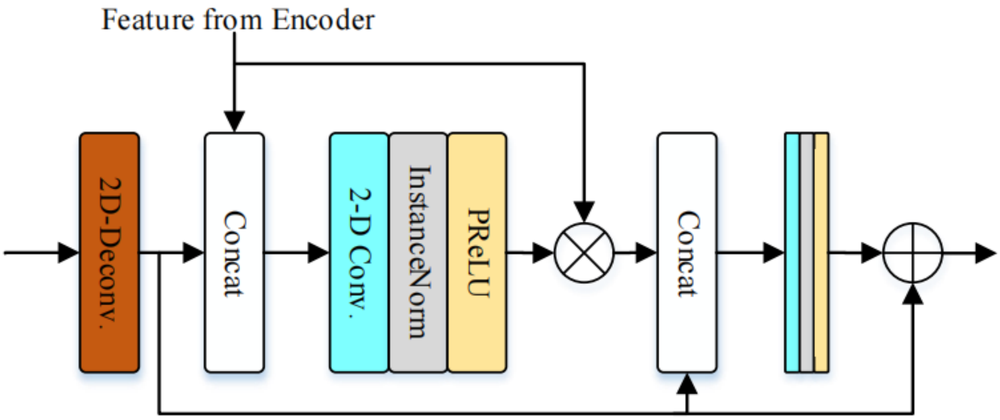

Figure 7. Detailed structure of the gated block inside the decoder.

As a supplement to spectrum estimation, the preliminary estimated spectrum is coupled with the noisy speech phase, Y_p, to obtain a roughly denoised complex spectrum diagram. Then, it is added element wise with the output ($\hat{X}'_r$,$\hat{X}'_i$) of the complex spectrum estimation decoder block to obtain the final complex spectrum diagram:

$$\hat{X}_r = \hat{X}'_m Y_m \cos Y_P + \hat{X}'_r \tag{6}$$

$$\hat{X}_i = \hat{X}'_m Y_m \sin Y_P + \hat{X}'_i \tag{7}$$

$$\hat{X}_m = \sqrt{\hat{X}_r{}^2 + \hat{X}_i{}^2} \tag{8}$$

where $\hat{X}_m$, $\hat{X}_r$ and $\hat{X}_i$ represent the magnitude, the real component of spectrum and the imaginary component of spectrum of the enhanced speech.

3.2. Two-Stage CGA Block

The two-stage CGA block consists of two cascaded CGAUs, namely, the time convolution-augmented gated attention unit (CGAU-T) and the frequency convolution-augmented gated attention unit (CGAU-F), which is shown in Figure 8. First, the input feature map, $D \in \mathbb{R}^{B \times T \times F' \times C}$, is reshaped to $D^T \in \mathbb{R}^{BF' \times T \times C}$ and input into CGAU-T to capture the time dependence. Then, the D_o^T and D^T are element-wise added and reshaped to $D^F \in \mathbb{R}^{BT \times F' \times C}$ and input into CGAU-F to capture the frequency dependence. Finally, output D_o^F and input D^F are element-wise added and reshaped to the final output, $D_o \in \mathbb{R}^{B \times T \times F' \times C}$.

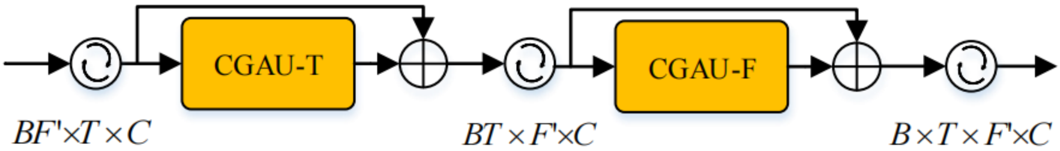

Figure 8. Two-stage CGA block architecture.

The CGAU-T and CGAU-F blocks have the same structure and different shaping operations. The structure of CGAU is shown in Figure 9 and is composed of a convolution block and a GAU. The input is connected to the output by a residual connection. We use the same structure as the convolution block in the conformer. The convolution block starts with a layer normalization. After that, the feature map is fed into a gating mechanism composed of a point-wise convolution, followed by GLU. Then, the output of the GLU is fed into a depth-wise convolution layer and activated by the swish function. Finally, a point-wise convolution layer restores the channel number.

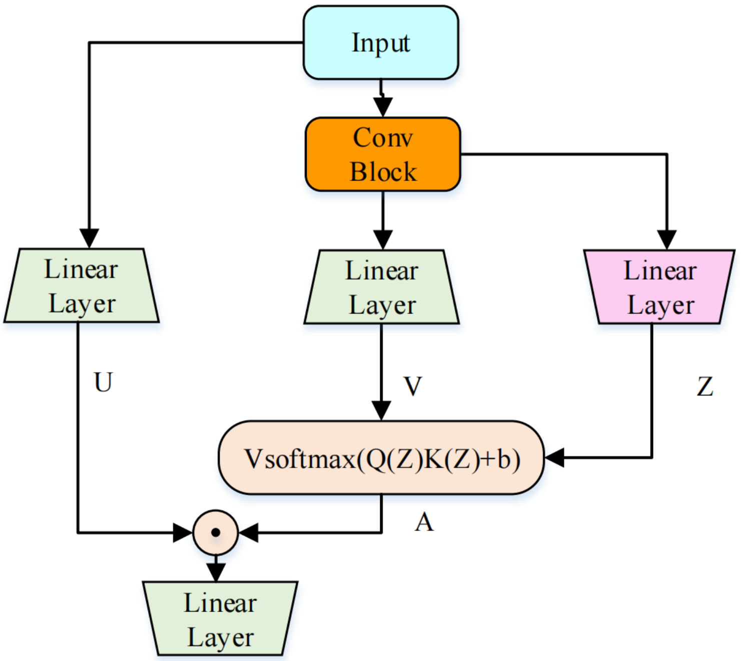

Figure 9. Our proposed convolution-augmented gated attention unit.

Taking CGAU-T as an example, input D^T is divided into two feeds, one of which is fed into the convolution Block. The query, key, and value are all replicas of the convolution block output, $D_{co}^T \in \mathbb{R}^{BF' \times T \times C}$. Then, the scaled dot-product attention is applied to the query, key, and value afterward as

$$Z = \varnothing_z\left(D_{co}^T W_z\right) \tag{9}$$

$$V = \varnothing_v\left(D_{co}^T W_v\right) \tag{10}$$

$$A = softmax\left(\frac{\mathcal{Q}(Z)\mathcal{K}(Z)^\intercal + b}{\sqrt{d}}\right)V \tag{11}$$

where $\varnothing$ represents the swish activation function, and W_z and W_v represent the learnable parameter matrixes of the linear layers. Z is a shared representation; $\mathcal{Q}$ and $\mathcal{K}$ are simple affine transformations that apply per-dim scalars and offsets to Z to obtain the query and the key. V represents values in the self-attention mechanism. b represents the rotation position coding [32]. d represents the dimension.

The other feed of input D^T passes through the linear layer and is activated by swish to obtain U. Finally, the Hadamard product of U and A is calculated and input into the linear layer to obtain output D_o^T so that the convolution-augmented attention information is introduced to the gated linear unit. The calculation formula is as follows:

$$U = \varnothing_u\left(D^T W_u\right) \tag{12}$$

$$D_o^T = (U \circ A)W_o \tag{13}$$

where $\circ$ represents Hadamard product. In addition, our proposed CGAU uses softmax as the activation function instead of the ReLU^2 used in GAU.

3.3. Metric Discriminator

The metric discriminator can mimic the metric score, which is nondifferentiable, so that it can be used as the loss function. In this paper, we use the normalized PESQ as the metric score. As shown in Figure 10, the metric discriminator consists of four Conv2D layers. The channels are 32, 64, 128, and 256. After four Conv2D layers, there is a global average pooling to handle the variable-length input. Finally, there are two linear layers and one sigmoid activation. When training the discriminator, we take both inputs as clean magnitudes to estimate the maximum metric score. Then, we take the inputs as the clean magnitudes and the enhanced magnitudes to estimate the metric score of the enhanced speech to approach their corresponding PESQ label. In addition, the trained generator can render enhanced speech approaching a normalized PESQ label of one.

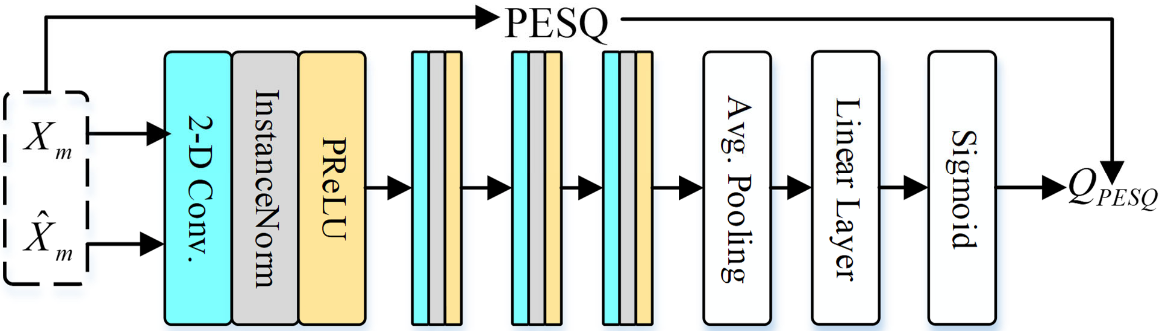

Figure 10. Detailed structure of the metric discriminator.

3.4. Loss Function

The loss function of the generator includes three terms:

$$L_G = \alpha L_{TF} + \beta L_{GAN} + \gamma L_{time} \tag{14}$$

where α, β, and γ are the weighting coefficients of the three loss terms in the total loss. Here, we take α as 1, β as 0.05, and γ as 0.2. L_{TF} represents the combination of magnitude loss, L_{Mag}, and phase-aware loss, L_{RI}:

$$L_{Mag} = E_{X_m,\hat{X}_m}\left[\|X_m - \hat{X}_m\|^2\right] \tag{15}$$

$$L_{RI} = E_{X_r,\hat{X}_r}\left[\|X_r - \hat{X}_r\|^2\right] + E_{X_i,\hat{X}_i}\left[\|X_i - \hat{X}_i\|^2\right] \tag{16}$$

$$L_{TF} = mL_{Mag} + (1-m)L_{RI} \tag{17}$$

where m represents the chosen weight, and we take $m = 0.7$. L_{GAN} represents the adversarial loss. The expression of L_{GAN} is

$$L_{GAN} = E_{X_m,\hat{X}_m}\left[\|D(X_m - \hat{X}_m) - 1\|^2\right] \tag{18}$$

where D represents the discriminator. Correspondingly, the expression of the adversarial loss in the discriminator is

$$L_D = E_{X_m}\left[\|D(X_m - X_m) - 1\|^2\right] + E_{X_m,\hat{X}_m}\left[\|D(X_m - \hat{X}_m) - Q_{PESQ}\|^2\right] \tag{19}$$

where Q_{PESQ} is the normalized PESQ score, and the value range is [0, 1]. In addition, some studies show that adding time loss, L_{time}, can improve the enhancement of speech [20]. The expression of L_{time} is

$$L_{time} = E_{x,\hat{x}}[\|x - \hat{x}\|_1] \tag{20}$$

4. Experiments

In this section, we first introduce the composition of the Voice Bank + DEMAND dataset for network training; then, we introduce the network training settings and six evaluation indicators for voice enhancement.

4.1. Datasets and Settings

The publicly available Voice Bank + DEMAND dataset was chosen to test our model. The speech database was obtained from the CSTR VCTK Corpus. The background noise database was obtained from the DEMAND database. The training set includes 11,572 sentences provided by 28 speakers, and the test set includes 824 sentences provided by 2 unseen speakers. We use eight natural and two artificial background noise processes to generate the training set under different SNR levels, ranging from 0 to 15 dB with an interval of 5 dB. We use five unseen background noise processes to generate the test set under different SNR levels, ranging from 2.5 to 17.5 dB with an interval of 5 dB.

All sentences are resampled to 16 kHz. For the training set, we slice the sentences into 2 s units, but there is no slicing in the test set. We use a Hamming window of length 25 ms and a hop length of 6.25 ms. Since we apply a power law compression with a compression coefficient of 0.3 to the spectrum [33] after STFT, it is reversed on the final estimated complex spectrum. Finally, we apply the inverse STFT to recover the time-domain signal. In addition, we use the AdamW optimizer to train both generator and discriminator for 100 epochs. The batch size is set to four. The learning rate of the discriminator is set to 0.001, which is twice that of the generator. After every 30 epochs, both learning rates will be halved.

4.2. Evaluation Indicators

We use six objective measures to evaluate the quality of the enhanced speech. For all metrics, higher scores indicate better performance.

PESQ [34]: Ranges from -0.5 to 4.5;

CSIG [35]: Mean opinion score (MOS) prediction of the signal distortion; ranges from 1 to 5;

CBAK [35]: MOS prediction of the background noise intrusiveness; ranges from 1 to 5;

COVL [35]: MOS prediction of the overall effect; ranges from 1 to 5;

SSNR: The segmental signal-to-noise ratio; ranges from 0 to ∞;

STOI [36]: The short-time objective intelligibility; ranges from 0 to 1.

5. Results and Discussion

In this section, we first conduct a comparison with baselines to verify the performance of the proposed model. Then, the effectiveness of CGA-MGAN is verified using ablation experiments. The experimental results are discussed and explained.

5.1. Baselines and Results Analysis

The CGA-MGAN we propose can be compared with some existing models objectively. These models include classical models, such as SEGAN [37], DCCRN [13], and Conv-TasNet [20], and state-of-the-art (SOTA) baselines. SEGAN is the earliest application of GAN in speech enhancement, and its network training is conducted by directly inputting time domain waveforms. DCCRN builds a complex convolution recursive network to simultaneously recover the magnitude and phase information in the TF domain by estimating the CRM. Conv-TasNet is a fully convolutional end-to-end time-domain speech separation network that can also be used for speech enhancement.

For methods based on the generation model, we choose three baselines, including TDCGAN [38], MetricGAN+ [39], UNIVERSE [40], and CDiffuSE [41]. TDCGAN first introduced dilated convolution and deep separable convolution to the GAN network and can build a time-domain speech enhancement system. It dramatically reduces network parameters and obtains a better speech enhancement effect than SEGAN. MetricGAN+ is an improved MetricGAN network for speech enhancement. It not only uses enhanced and clean speech to train the discriminator, but also uses noisy speech to minimize the distance between the discriminator and target objective metrics. Moreover, MetricGAN+'s generator uses the learnable sigmoid function for mask estimation, which improves the generator's speech enhancement ability. UNIVERSE builds a generative score-based diffusion model and a multi-resolution conditioning network so that mixture density networks can be enhanced and achieve good speech enhancement results. CDiffuSE learns the characteristics of noise signals and incorporates them into diffusion and reverse processes, making the model highly robust against changes in noise characteristics.

For methods based on the improved transformer, we choose four baselines, including SE-Conformer [25], DB-AIAT [17], DPT-FSNet [42], and DBT-Net [43]. SE-Conformer first introduced the convolution-augmented transformer to the field of speech enhancement. The proposed speech enhancement architecture can focus on the whole sequence through MHSA and convolutional neural networks to capture short-term and long-term time series information. DB-AIAT constructs a dual-branch network through the decoupling-style phase-aware method. It first estimates the magnitude spectrum roughly, and then the spectral details that the magnitude-marking branch missed are compensated. DPT-FSNET integrates subband band and complete band information and proposes a transformer-based dual-branch frequency domain speech enhancement network. DBT-Net proposes a dual-branch network to estimate magnitude and phase information. Interaction modules are introduced to obtain features learned from one branch to facilitate the information flow between branches to curb the undesired parts.

As shown in Table 1, compared with the improvement work of MetricGAN (MetricGAN+), there is a 0.32 improvement in the PESQ score. In addition, compared with the advanced generation model currently used in speech enhancement, CGA-MGAN achieves better performance. Finally, compared with recent methods based on improved transformers, such as DPT-FSNet, our method is also better in almost all evaluation scores. At the same time, the model size is relatively small (1.14 M).

Table 1. Performance comparison of Voice Bank + DEMAND dataset.

Method	Size(M)	PESQ	CSIG	CBAK	COVL	SSNR	STOI
Noisy	- *	1.97	3.35	2.44	2.63	1.68	0.91
SEGAN	97.47	2.16	3.48	2.94	2.80	7.73	0.92
DCCRN	3.70	2.68	3.88	3.18	3.27	-	0.94
Conv-TasNet	5.1	2.84	2.33	2.62	2.51	-	-
TDCGAN	5.12	2.87	4.17	3.46	3.53	9.82	0.95
MetricGAN+	-	3.15	4.14	3.16	3.64	-	-
UNIVERSE	-	3.33	-	-	3.82	-	0.95
CDiffuSE	-	2.52	3.72	2.91	3.10	-	-
SE-Conformer	-	3.13	4.45	3.55	3.82	-	0.95
DB-AIAT	2.81	3.31	**4.61**	3.75	3.96	10.79	0.96
DPT-FSNet	0.91	3.33	4.58	3.72	4.00	-	0.96
DBT-Net	2.91	3.30	4.59	3.75	3.92	-	0.96
CGA-MGAN	1.14	**3.47**	4.56	**3.86**	**4.06**	**11.09**	**0.96**

* "-" denotes that the result was not provided in the original paper.

5.2. Ablation Study

To investigate the contribution of the different CGA-MGAN components proposed to enhance performance, we have conducted an ablation study. Several variants of the CGA-MGAN model are compared in Table 2: (i) removing the convolution block in CGAU; (that is, using GAU to replace CGAU (w/o Conv. Block)); (ii) using conformer to replace CGAU in the two-stage block (using a conformer); (iii) removing gating blocks in the decoder (w/o gating decoders); (iv) removing the phase compensation mechanism and only improving the speech in the magnitude spectrum (Mag-only); (v) removing the discriminator (w/o discriminator).

Table 2. Results of the ablation study.

Method	PESQ	CSIG	CBAK	COVL	SSNR	STOI
CGA-MGAN	**3.47**	**4.56**	**3.86**	**4.06**	**11.09**	**0.96**
w/o Conv. Block	3.37	4.50	3.80	3.97	10.97	0.96
Using Conformer	3.33	4.43	3.72	3.91	10.18	0.96
w/o Gating Decoders	3.43	4.52	3.83	4.02	10.99	0.96
Mag-only	3.42	4.54	3.80	4.03	10.73	0.96
w/o Discriminator	3.37	4.54	3.79	4.00	10.83	0.96

We set all variants using the same configuration as CGA-MGAN. As shown in Table 2, all these variants underperform the proposed CGA-MGAN. Comparing CGA-MGAN with (i), a decrease of 0.1 in PESQ can be observed because GAU cannot extract the short-term features of speech very well after the convolution block is removed. Comparing CGA-MGAN with (ii), a decrease of 0.14 in PESQ can be observed. This proves that the CGAU model we propose is superior to traditional convolution-augmented transformers for speech enhancement. Comparing CGA-MGAN with (iii) and (iv), we find that gating blocks and the phase-aware method are essential to CGA-MGAN. Because the gating block can retain important features in the encoder and suppresses irrelevant features, it can also control the information flow of the network and simulate complex interactions, which is quite effective for improving the network's speech enhancement ability [22]. In addition, the decoupling-style phase-aware method can process the coarse-grained regions and fine-grained regions of the spectrum in parallel so that lost spectrum details can be compensated for, and it can avoid the compensation effect between magnitude and phase [17]. Finally, comparing CGA-MGAN with (v), we can observe that removing the discriminator has a negative impact on all given scores, which proves the advantages of using MetricGAN.

In addition, we can focus on CGA-MGAN and (ii) and use real-time factors (RTFs) to compare their computational complexity. RTFs can be measured by taking an average of five runs on an Intel Xeon Silver 4210 CPU. The RTF of CGA-MGAN is 0.017, while the RTF of (ii) is 0.025. This proves that the CGAU we propose has lower computational complexity than traditional convolution-augmented transformers.

6. Conclusions

In this work, we propose CGA-MGAN for speech enhancement, which can better combine the advantages of convolution and self-attention. Our approach combines CGAU, which can capture time and frequency dependencies with lower computational complexity and achieve better results, together with an encoder–decoder structure, including gating blocks using the decoupling-style phase-aware method, which can collaboratively estimate the magnitude and phase information of clean speech and avoid the compensation effect. Experiments on Voice Bank + DEMAND show that the CGA-MGAN model we propose performs excellently at a relatively small model size (1.14 M). In the future, we will further study the application of CGAU to other speech tasks, such as separation and dereverberation.

Author Contributions: Conceptualization, H.C.; funding acquisition, X.Z.; methodology, H.C.; project administration, X.Z.; software, H.C.; validation, H.C.; writing—original draft, H.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (No: 61172017).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: A publicly available dataset (Voice Bank + DEMAND) was analyzed in this study. The Voice Bank + DEMAND dataset (accessed on 17 December 2021) can be found here: https://datashare.ed.ac.uk/handle/10283/2791.

Acknowledgments: The authors would like to thank all the anonymous reviewers for their helpful comments and suggestions to improve the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, D.; Chen, J. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 1702–1726. [CrossRef]
2. Atmaja, B.T.; Farid, M.N.; Arifianto, D. Speech enhancement on smartphone voice recording. In *Journal of Physics: Conference Series*; IOP Publishing: Bristol, UK, 2016; p. 012072.
3. Tasell, D.J.V. Hearing loss, speech, and hearing aids. *J. Speech Lang. Hear. Res.* **1993**, *36*, 228–244. [CrossRef] [PubMed]
4. Wang, J.; Liu, H.; Zheng, C.; Li, X. Spectral subtraction based on two-stage spectral estimation and modified cepstrum thresholding. *Appl. Acoust.* **2013**, *74*, 450–458. [CrossRef]
5. Abd El-Fattah, M.A.; Dessouky, M.I.; Abbas, A.M.; Diab, S.M.; El-Rabaie, E.-S.M.; Al-Nuaimy, W.; Alshebeili, S.A.; Abd El-samie, F.E. Speech enhancement with an adaptive Wiener filter. *Int. J. Speech Technol.* **2014**, *17*, 53–64. [CrossRef]
6. Ephraim, Y.; Malah, D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.* **1984**, *32*, 1109–1121. [CrossRef]
7. Wang, D. On ideal binary mask as the computational goal of auditory scene analysis. In *Speech Separation by Humans and Machines*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 181–197.
8. Narayanan, A.; Wang, D. Ideal Ratio Mask Estimation using Deep Neural Networks for Robust Speech Recognition. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 7092–7096.
9. Wang, Y.; Narayanan, A.; Wang, D. On training targets for supervised speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 1849–1858. [CrossRef]
10. Xu, Y.; Du, J.; Dai, L.-R.; Lee, C.-H. A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *23*, 7–19. [CrossRef]
11. Paliwal, K.; Wójcicki, K.; Shannon, B. The importance of phase in speech enhancement. *Speech Commun.* **2011**, *53*, 465–494. [CrossRef]
12. Yin, D.; Luo, C.; Xiong, Z.; Zeng, W. Phasen: A Phase-And-Harmonics-Aware Speech Enhancement Network. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 9458–9465.
13. Hu, Y.; Liu, Y.; Lv, S.; Xing, M.; Zhang, S.; Fu, Y.; Wu, J.; Zhang, B.; Xie, L. DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. *arXiv* **2020**, arXiv:2008.00264.
14. Williamson, D.S.; Wang, Y.; Wang, D. Complex ratio masking for monaural speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *24*, 483–492. [CrossRef]
15. Li, A.; Zheng, C.; Yu, G.; Cai, J.; Li, X. Filtering and Refining: A Collaborative-Style Framework for Single-Channel Speech Enhancement. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 2156–2172. [CrossRef]
16. Wang, Z.-Q.; Wichern, G.; Le Roux, J. On the compensation between magnitude and phase in speech separation. *IEEE Signal Process. Lett.* **2021**, *28*, 2018–2022. [CrossRef]
17. Yu, G.; Li, A.; Zheng, C.; Guo, Y.; Wang, Y.; Wang, H. Dual-Branch Attention-In-Attention Transformer for Single-Channel Speech Enhancement. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 7847–7851.
18. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
19. Subakan, C.; Ravanelli, M.; Cornell, S.; Bronzi, M.; Zhong, J. Attention is All You Need in Speech Separation. In Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 21–25.
20. Luo, Y.; Mesgarani, N. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1256–1266. [CrossRef]
21. Pandey, A.; Wang, D. TCNN: Temporal Convolutional Neural Network for Real-Time Speech Enhancement in the Time Domain. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6875–6879.
22. Jia, X.; Li, D. TFCN: Temporal-Frequential Convolutional Network for Single-Channel Speech Enhancement. *arXiv* **2022**, arXiv:2201.00480.
23. Gulati, A.; Qin, J.; Chiu, C.-C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y. Conformer: Convolution-augmented transformer for speech recognition. *arXiv* **2020**, arXiv:2005.08100.
24. Chen, S.; Wu, Y.; Chen, Z.; Wu, J.; Li, J.; Yoshioka, T.; Wang, C.; Liu, S.; Zhou, M. Continuous Speech Separation with Conformer. In Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 5749–5753.

25. Kim, E.; Seo, H. SE-Conformer: Time-Domain Speech Enhancement Using Conformer. In Proceedings of the Interspeech, Brno, Czech Republic, 30 August–3 September 2021; pp. 2736–2740.
26. Cao, R.; Abdulatif, S.; Yang, B. CMGAN: Conformer-based Metric GAN for Speech Enhancement. *arXiv* **2022**, arXiv:2203.15149.
27. Fu, S.-W.; Liao, C.-F.; Tsao, Y.; Lin, S.-D. Metricgan: Generative Adversarial Networks Based Black-Box Metric Scores Optimization for Speech Enhancement. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 2031–2041.
28. Valentini-Botinhao, C.; Wang, X.; Takaki, S.; Yamagishi, J. Investigating RNN-Based Speech Enhancement Methods for Noise-Robust Text-to-Speech. In Proceedings of the SSW, Sunnyvale, CA, USA, 13–15 September 2016; pp. 146–152.
29. Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.; Wang, Z.; Smolley, S.P. Least Squares Generative Adversarial Networks. In Proceedings of the Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2794–2802.
30. Hua, W.; Dai, Z.; Liu, H.; Le, Q. Transformer Quality in Linear Time. In Proceedings of the International Conference on Machine Learning, Guangzhou, China, 18–21 February 2022; pp. 9099–9117.
31. Lu, Y.; Li, Z.; He, D.; Sun, Z.; Dong, B.; Qin, T.; Wang, L.; Liu, T.-Y. Understanding and improving transformer from a multi-particle dynamic system point of view. *arXiv* **2019**, arXiv:1906.02762.
32. Su, J.; Lu, Y.; Pan, S.; Wen, B.; Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *arXiv* **2021**, arXiv:2104.09864.
33. Braun, S.; Tashev, I. A Consolidated View of Loss Functions for Supervised Deep Learning-Based Speech Enhancement. In Proceedings of the 2021 44th International Conference on Telecommunications and Signal Processing (TSP), Brno, Czech Republic, 26–28 July 2021; pp. 72–76.
34. Rix, A.W.; Beerends, J.G.; Hollier, M.P.; Hekstra, A.P. Perceptual Evaluation of Speech Quality (PESQ)—A New Method for Speech Quality Assessment of Telephone Networks and Codecs. In Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221), Salt Lake City, UT, USA, 7–11 May 2001; pp. 749–752.
35. Hu, Y.; Loizou, P.C. Evaluation of objective quality measures for speech enhancement. *IEEE Trans. Audio Speech Lang. Process.* **2007**, *16*, 229–238. [CrossRef]
36. Taal, C.H.; Hendriks, R.C.; Heusdens, R.; Jensen, J. A Short-Time Objective Intelligibility Measure for Time-Frequency Weighted Noisy Speech. In Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 14–19 March 2010; pp. 4214–4217.
37. Pascual, S.; Bonafonte, A.; Serra, J. SEGAN: Speech enhancement generative adversarial network. *arXiv* **2017**, arXiv:1703.09452.
38. Ye, S.; Hu, X.; Xu, X. Tdcgan: Temporal dilated convolutional generative adversarial network for end-to-end speech enhancement. *arXiv* **2020**, arXiv:2008.07787.
39. Fu, S.-W.; Yu, C.; Hsieh, T.-A.; Plantinga, P.; Ravanelli, M.; Lu, X.; Tsao, Y. Metricgan+: An improved version of metricgan for speech enhancement. *arXiv* **2021**, arXiv:2104.03538.
40. Serrà, J.; Pascual, S.; Pons, J.; Araz, R.O.; Scaini, D. Universal Speech Enhancement with Score-based Diffusion. *arXiv* **2022**, arXiv:2206.03065.
41. Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
42. Dang, F.; Chen, H.; Zhang, P. DPT-FSNet: Dual-Path Transformer Based Full-Band and Sub-Band Fusion Network for Speech Enhancement. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 6857–6861.
43. Yu, G.; Li, A.; Wang, H.; Wang, Y.; Ke, Y.; Zheng, C. DBT-Net: Dual-branch federative magnitude and phase estimation with attention-in-attention transformer for monaural speech enhancement. *arXiv* **2022**, arXiv:2202.07931. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.