Computer Science and Application 计算机科学与应用, 2026, 16(6), 90-102
Published Online June 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.166211

Hans汉斯

U-Net网络中结合卷积与注意力机制的跳跃连接的单通道语音增强

谭应伟

北京享来智商贸中心技术部，北京

收稿日期：2026年5月10日；录用日期：2026年6月15日；发布日期：2026年6月24日

摘　要

在可懂度和感知质量方面，单通道语音增强技术从深度学习的成功中获得了巨大收益。传统的方法侧重于应用U-Net模型来预测带噪语音的纯净信号，这种模型的跳跃连接以及序列建模模块存在局限性。本研究提出了在U-Net网络中结合卷积与注意力机制的跳跃连接的单通道语音增强算法。一方面，基于卷积的跳跃连接(convolution skip)应用含有卷积门控机制的卷积模块来提取更重要的局部特征信息；另一方面，基于注意力机制的跳跃连接(attention skip)结合了ROPE位置编码与图卷积网络(GCN)，从而能够更好地提取上下文全局特征信息；除此之外，conformer-block模块应用了卷积门控单元(CGU)与多头增强的注意力机制，它能够更好的建模序列信息。在VoiceBank-DEMAND语音数据集上对提出的方法进行了验证，在噪声数据上获得了0.6975的语音感知质量评估(PESQ)提升、0.0124的语音短时客观可懂度(STOI)提升以及8.4324的分段信噪比(SSNR)提升。实验结果表明，与基线denoiser方法相比，提出来的方法更有优越性。

关键词

单通道语音增强，卷积门控单元，图卷积网络

Combining Convolution and Attention with Skip Connection in U-Net for Monaural Speech Enhancement

Yingwei Tan

Technical Department, Beijing Xianglaizhi Commerce and Trade Center, BeiJing

Received: May 10, 2026; accepted: June 15, 2026; published: June 24, 2026

文章引用: 谭应伟. U-Net 网络中结合卷积与注意力机制的跳跃连接的单通道语音增强[J]. 计算机科学与应用, 2026, 16(6): 90-102. DOI: 10.12677/csa.2026.166211

Abstract

In terms of intelligibility and perceptual quality, single-channel speech enhancement technology has benefited greatly from the success of deep learning. Traditional methods focus on applying the U-Net model to predict the clean signal from noisy speech, but this model's skip connections and sequence modeling modules have limitations. This study proposes a single-channel speech enhancement algorithm that combines convolution and attention mechanisms in the skip connections of the U-Net network. On the one hand, the convolution skip connection applies a convolution module with a convolution gating mechanism to extract more important local feature information. On the other hand, the attention skip connection combines ROPE position encoding and graph convolutional networks (GCN) to better extract contextual global feature information. Besides, the conformer-block module applies a convolutional gated unit (CGU) and an enhanced multi-head attention mechanism, which can better model sequence information. The proposed method was validated on the VoiceBank-DEMAND speech dataset, achieving an improvement of 0.6975 in Perceptual Evaluation of Speech Quality (PESQ), 0.0124 in Short-Term Objective Intelligibility (STOI), and 8.4324 in Segmental Signal-to-Noise Ratio (SSNR) on noisy data. Experimental results show that the proposed method is superior to the baseline denoiser method.

Keywords

Monaural Speech Enhancement, Convolutional Gated Unit, Graph Convolutional Network

1. 引言

在真实的环境中，感知到的语音质量与可懂度直接依赖于底层语音增强系统的性能。因此，语音增强框架也是现代语音识别系统中不可或缺的一部分。关于单通道的语音增强，许多算法已经被提了出来。谱减法[1]、基于统计模型的方法[2]、维纳滤波法[3]以及子空间法[4] [5]都是经典的语音增强算法。尽管这些方法在处理平稳噪声时表现良好，但在低信噪比或者非平稳噪声情况下，单通道语音增强依然存在巨大的问题。声学环境是极其复杂的。驾驶过程中会受到来自车内和车外的各种噪音干扰，如风声、引擎声、车轮声、背景音乐以及其他说话人声的干扰。这些类型噪声的存在会大大降低人类语音的可懂度。最近，基于深度神经网络(DNN)的模型[6]-[12]在处理平稳和非平稳噪声方面表现显著优于传统方法，同时在客观和主观评估中生成了更高质量的语音。

按照监督学习问题的思路，深度神经网络可以在时频域或直接在时域中增强带噪的语音。时频域方法应用于频谱图之上，其依据是短时傅里叶变换(STFT)后，语音和噪声的精细结构在时频表示下更易于分离。最近，在自动语音识别(ASR)和语音分离任务中，由于 conformer 能够同时捕捉局部和全局上下文的信息，因此已被提出作为 transformer 的替代方案[13]-[15]。在文献[16]中，conformer 的结构被继承到了 Conv-TasNet 的框架中。根据文献[7]，它引入了两阶段的 conformer 模块，其能够以相对较低的计算复杂度捕捉时间和频率依赖性。该方法采用了一种度量判别器，有助于在不对其他度量产生不利影响的情况下，提高相应的评估度量。卷积循环网络(CRN)采用卷积的编码器－解码器结构，能够有效提取高级特征，以便更好地从嘈杂的语音频谱图中分离出干净信息。但是，卷积循环网络仅利用了幅度分量，而忽

略了相位。因为相位分量中的随机结构，这导致了它很难被其他方法所应用[17] [18]。为了绕过相位估计问题，近期的方法采用了增强复数频谱图的策略，这隐式地增强了幅度和相位。深度复数卷积循环网络模型(DCCRN) [19]利用了一个复杂的网络结构来对复数频谱进行建模，并优化了尺度不变的信噪比(SI-SNR)损失。与 DCCRN 类似，文献[10]以 U-Net 为骨架，使用混合 Encoder 和 Decoder 架构来同时对复数谱和幅度谱进行建模，并使用双路注意力机制来分别进行局部时序和频带建模。这样显著提升了增强性能。

面向时域的方法[20]-[22]是基于生成模型的，这些模型经过训练，能够直接从噪声波形中估计出干净波形的片段。在这种情况下，波形中的全部信息得以保留，无需任何转换或重建要求。语音增强生成对抗网络(SEGAN) [6]提供了一种快速的增强过程。无需因果关系。它以原始音频为输入，进行端到端的处理。因此，不会提取任何手工特征。文献[20]提出了一个作用于原始波形的因果语音增强模型(denoiser)。该去噪模型基于带有跳跃连接的编码器–解码器架构，并采用 DEMUCS [23]架构进行语音增强。这是一种在波形领域为音乐源分离而开发的先进算法。作者将其改造成了一种因果语音增强器。在文献[20]的基础上，文献[21]提出了一种端到端的语音恢复方法，HDDEMUCS 模型利用两个异构解码器从抑制和细化两个不同角度进行语音恢复。它通过抑制解码器的掩码估计方法表现出了强大的抑制能力，并且展示出了带扩张卷积层的细化解码器的有效性。此外，它还加入了一个融合模块，通过预测一个可学习的加权值来有效地结合两个解码器的输出。在文献[24]中，设计了一个适用于端到端语音增强的改进的 U-Net 网络模型。与基线 U-Net 网络相比，一方面通过加入空洞卷积减小由采样带来的信息损失；另一方面引入了注意力机制结构，结合了含噪语音更多的上下文信息，提取更深层次和更丰富的特征信息。虽然文献[24]利用膨胀卷积扩大了感受野，改善了语音增强的性能，但是它对长时带噪语音中的语音成分关注不够。在文献[25]中，提出一种融合多头注意力机制和 U-Net 深度网络的增强模型 TU-net，实现基于时域的端到端单通道语音增强。TU-net 网络模型采用 U-Net 网络的编解码层对带噪语音信号进行多尺度特征融合，并利用多头注意力机制实现双路径 Transformer，用于计算语音掩模，更好地建模长时相关性。在文献[9]中，在传统 U-Net 基础上，引入多层嵌套的残差块，形成更深层次的特征提取路径。在每一层嵌套结构中，设计两个层级的跳跃连接，以融合不同尺度的上下文信息，缓解语义鸿沟。在每个残差块输出端加入因果时间频率注意力模块，动态建模多尺度语音上下文，提升增强效果。

在本研究中，提出了一个 U-Net 网络中结合卷积与注意力机制的跳跃连接的单通道语音增强算法。主要贡献在以下三个方面。第一，卷积跳跃连接是基于卷积门控单元(CGU)的，它帮助从卷积模块中自适应筛选重要局部信息。同时，它缓解了编码器与解码器之间的信息鸿沟，减少了信息冗余。第二，多头注意力跳跃连接是基于 RoPE 位置编码及图卷积网络的。它能够学习多层次的上下文依赖关系，也能消除信息鸿沟及冗余。RoPE [26]是一种位置嵌入方法，它利用旋转矩阵来编码绝对位置信息，并在自注意力公式中自然地融入了明确的相对位置依赖关系。值得注意的是，RoPE 具有几个有价值的特性，包括可以灵活地扩展到任何序列长度，以及随着相对距离的增加，token 间的依赖关系会衰减。图卷积网络模型(GCN) [27]是一种神经网络架构，能够利用图结构并以卷积的方式从邻域中聚合节点信息。我们提出利用图神经网络来捕捉多头模块中的头相关性。它允许在注意力头之间传递信息。第三，编码器与解码器架构的嵌入空间被 conformer-block 建模。它同样基于卷积门控单元与多头增强的注意力机制，能够建模全局与局部的上下文信息，增强了对长时语音中语音成分的关注。

本研究采用语音质量感知评估(Perceptual Evaluation of Speech Quality, PESQ) [28]、短时客观可懂度(Short-Time Objective Intelligibility, STOI) [29]及分段信噪比(Segmental Signal-to-Noise Ratio, SSNR) [30]作为核心评估指标，系统量化所提算法的语音增强性能。在 VoiceBank-DEMAND 公开语料库上开展对比实验，结果表明，相较于传统降噪方法 denoiser，本研究提出的算法在各项指标上均实现了性能提升，验证

了其在复杂噪声环境下的有效性。

2. 单通道语音增强方案设计

U-Net 网络中结合卷积与注意力机制的跳跃连接的单通道语音增强框架如图 1 所示，该框架由一个 U-Net 网络模块和一个 conformer-block 序列建模模块组成。它在 U-Net 网络的编码器－解码器的嵌入底层中加入了 conformer-block 序列模型，并且在跳跃连接中采用了卷积或注意力跳跃连接。这样不仅可以减少编码器与解码器之间的语义鸿沟及信息冗余，而且可以利用 conformer-block 的长时相关性预测和 U-Net 的多尺度特征融合优势，提高语音的质量和可懂度。由于参数量过大的原因，注意力跳跃连接(attention skip)只在跳跃连接的倒数第二层使用。

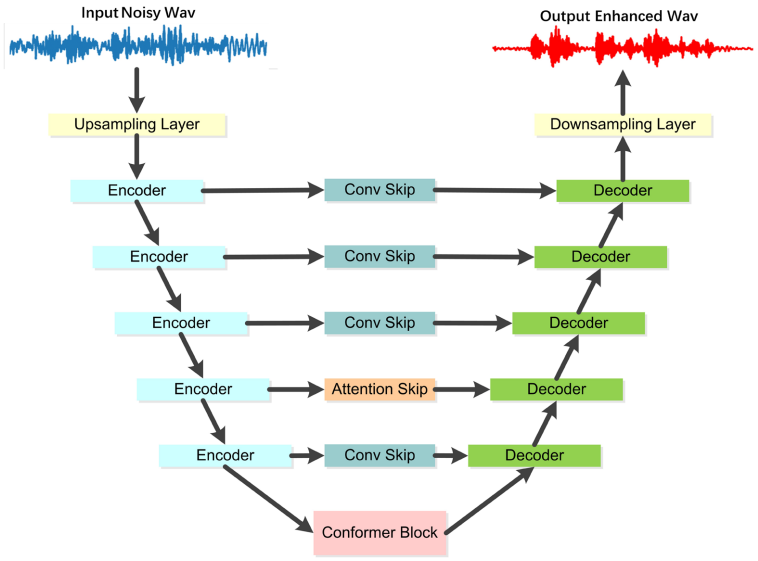

Figure 1. The monaural speech enhancement framework based on combining convolution and attention with skip connection in U-Net.
图 1. U-Net 网络中结合卷积与注意力机制的跳跃连接的单通道语音增强框架图

U-Net 网络模块由编码模块和解码模块组成，并使用卷积或注意力跳跃连接结构将编码器与解码器相连。U-Net 将编码器提取到的特征通过卷积或注意力跳跃连接传递到解码模块，实现了不同尺度下的特征融合。在实际应用中，浅层特征包含更多的语音细节信息，而深层特征包含更多的上下文包络信息。通过融合不同尺度的特征信息有助于消除噪声信号。

2.1. 跳跃连接

跳跃连接是神经网络架构中常用的一种技术，旨在促进信息流动和梯度传播[31]。传统的 U-Net 模型采用跳跃连接机制，其核心思想是将编码器的特征图与解码器对应层的特征图连接起来，以提供更多高级语义信息。然而，这种机制在实践中可能会导致特定问题，如信息瓶颈和梯度消失。为了解决这些问题，本文提出了一种改进的跳跃连接机制。

在本文的模型中，采用了一种改进的跳跃连接机制来解决传统 U-Net 中简单拼接操作可能导致的信息传递限制问题。在跳跃连接过程中，利用卷积或注意力机制来调节编码器和解码器之间的特征。具体而言，一方面，利用含卷积门控单元的卷积模块来筛选重要特征信息并加强局部特征学习；另一方面，

使用注意力机制来动态调整连接权重，使模型能够更加关注重要的特征信息，同时抑制无关信息。这种跳跃连接机制与注意力机制相结合，使模型能够更灵活地利用编码器和解码器之间的特征交互，从而提高模型在语音增强中的准确性和鲁棒性。

2.2. 卷积跳跃连接(Conv Skip)

卷积跳跃连接的结构如图 2 所示。它由一个逐点卷积和卷积门控单元开始，逐点卷积将通道从 d 映射到 $2d$，为 CGU 准备两路数据。然后，进行深度卷积扩大感受野。接下来，进行 Batchnorm，Swish 以及逐点卷积等操作。最后，完成 Dropout 操作。

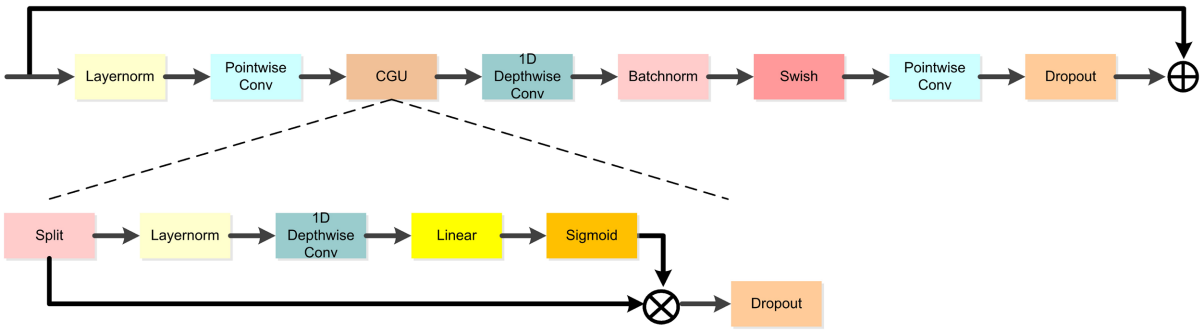

Figure 2. Convolutional skip connection structure based on convolutional gated unit
图 2. 基于卷积门控单元的卷积跳跃连接结构图

其中，关键模块卷积门控单元(CGU)应用了一个深度卷积模块来捕获了局部依赖。首先输入特征序列 $Z \in \mathbb{R}^{T\times 2d}$ 被沿着特征维度进行相等切分，得到两个新的序列 $Z_1, Z_2 \in \mathbb{R}^{T\times d}$。接下来，$Z_2$ 进行 Layernorm 处理，深度卷积处理，线性变换处理以及 Sigmoid 激活处理，公式如下：

$$Z_2' = \text{Sigmoid}\left(\text{FC}\left(\text{DWConv}\left(\text{Layernorm}\left(Z_2\right)\right)\right)\right) \tag{1}$$

CGU 模块的最后输出是逐元素点积 $\tilde{Z} = Z_1 \odot Z_2'$。这个门控机制解决了选择重要特征信息的问题。

2.3. 注意力跳跃连接(Attention Skip)

2.3.1. 图卷积网络

图神经网络(GNNs)是对传统神经网络的重要扩展，专门设计用于处理编码为图形的非欧几里得数据结构。在图神经网络的众多子类型中，卷积图神经网络因其遵循类似于卷积神经网络中采用的权重共享原则而脱颖而出。在本研究中，我们利用多层图卷积网络[32]来构建针对我们特定任务的图卷积神经网络。通过采用动态图构建方法，我们的框架能够以高度针对每个单独音频输入的方式有效捕捉多头信息，从而增强我们模型的表征能力。无向图 $G=(V,\xi)$ 被建立，V 表示图结点的集合，v_i 表示头，也即是图的结点。ξ 是图的边的集合，(v_i, v_j) 记作图的边。在图论中，图由其邻接矩阵 $A \in \mathbb{R}^{|V|\times|V|}$ 和度矩阵 D 来表征。这里考虑的是加权邻接矩阵，A 的元素相应于在两个结点之间一个带权重的边 $(v_i, v_j) \in \xi$。直观上，每个权重都反映了图中两个结点的特征向量之间的相似性。在本方法中，邻接矩阵权 A 的权重 w_{ij} $(i, j \in |V|)$ 在训练过程中学习得到。

图 G 提供了一个结构化的方式来捕获所有头结点之间的信息。现在可以利用图卷积网络(GCNs)从这张图中学习头部关系。应用 GCN 通过学习每个节点相对于其邻居的表示，来为节点特征学习更高的抽象层次。假定图 $G=(V,\xi)$，图卷积网络(GCN)对输入特征矩阵 $X \in \mathbb{R}^{|V|\times N}$ 进行非线性变换，N 是特征的维数。从数学角度来看，图卷积网络(GCN)可以表示为：

$$H^{(l)}=g\left(D^{-1/2}AD^{-1/2}H^{(l-1)}W^{(l-1)}\right) \tag{2}$$

其中，$H^{(l)}\in\mathbb{R}^{|V|\times K}$ 表示第 l 层的输出，它有 K 个特征，$H^{(0)}=X$。D 是一个对角线节点度矩阵，$W^{(l-1)}$ 是第 $l-1$ 层可学习的权重矩阵，g 是激活函数。

2.3.2. 旋转位置编码(RoPE)

按照文献[26]中的描述，从维数 $d=2$ 的基本状况开始，RoPE 有一个潜在的解决方案：

$$f_q\left(x_m,m\right)=\left(W_q x_m\right)\mathrm{e}^{im\theta} \tag{3}$$

$$f_k\left(x_n,n\right)=\left(W_k x_n\right)\mathrm{e}^{in\theta} \tag{4}$$

$$g\left(x_m,x_n,m-n\right)=\mathrm{Re}\left[\left(W_q x_m\right)\left(W_k x_n\right)^{*}\mathrm{e}^{i(m-n)\theta}\right] \tag{5}$$

其中，$\mathrm{Re}[\cdot]$表示复数取实部。$\left(W_k x_n\right)^{*}$表示$\left(W_k x_n\right)$的共轭复数。$\theta\in\mathbb{R}$ 表示一个非零的常数。

考虑到内积线性特性的优势，当 d 为偶数时，可以将解外推到任意维度。具体做法是，将 d 维空间划分为 $d/2$ 个子空间，然后将它们合并：

$$f_q\left(x_m,m\right)=R^{d}_{\Theta,m}W_q x_m \tag{6}$$

$$f_k\left(x_n,n\right)=R^{d}_{\Theta,n}W_k x_n \tag{7}$$

其中，

$$R^{d}_{\Theta,m}=\begin{pmatrix}X_1 & & & \\ & X_2 & & \\ & & \ddots & \\ & & & X_{d/2}\end{pmatrix} \tag{8}$$

$$\Theta=\left\{\theta_i=10000^{-2(i-1)/d},i\in\left[1,2,\cdots,d/2\right]\right\} \tag{9}$$

$$X_i=\begin{pmatrix}\cos m\theta_i & -\sin m\theta_i\\ \sin m\theta_i & \cos m\theta_i\end{pmatrix} \tag{10}$$

2.3.3. 基于 GCN 和 RoPE 的多头自注意力跳跃连接

图 3 详细展示了多头自注意力(MHSA)跳跃连接，它是基于图卷积神经网络(GCN)与旋转位置编码(RoPE)的。首先，GCN 被应用于中间特征上面，产生 x_n 作为输出。随后，位置信息被无缝地整合到输入序列中，从而使这些输入能够分别转换为查询向量、键向量和值向量。在 RoPE 的辅助下，位置编码的这种整合增强了模型的上下文理解能力，进而提升了自注意力能力。

$$q_n=f_q\left(x_n,n\right) \tag{11}$$

$$k_n=f_k\left(x_n,n\right) \tag{12}$$

$$v_n=x_n \tag{13}$$

其中，q_n,k_n 分别由函数 $f_q(\cdot),f_k(\cdot)$ 合并了第 n 个位置信息。最后，使用查询向量和键向量计算注意力权重，输出是值向量的加权和：

$$a_{n,n}=\mathrm{softmax}\left(\frac{q_n k_n^{\mathrm{T}}}{\sqrt{d}}\right) \tag{14}$$

$$o_n=\sum_{n=1}^{T}a_{n,n}v_n \tag{15}$$

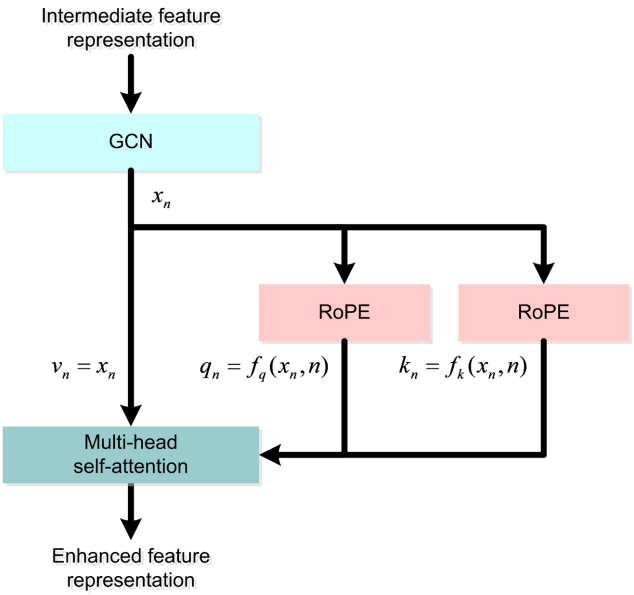

Figure 3. Multi-head self attention skip connection structure based on GCN and RoPE
图 3. 基于 GCN 和 RoPE 的多头自注意力跳跃连接结构图

2.4. Conformer 块(Conformer block)

图 4 描述了 Conformer Block 的结构图。具体构造方式为由两个 Feed Forward Module 夹着一个多头自注意力模块(Multi-Head Self-Attention Module)和一个卷积模块(Convolution Module)的类三明治结构，每个模块都有各自的残差模块，其中 Feed Forward Module 中残差系数设为 1/2。其中，多头自注意力模块采用了 RoPE 位置编码，并进行了 GCN 的多头增强处理；卷积模块应用了卷积门控单元(CGU)进行局部特征提取和重要信息选择。

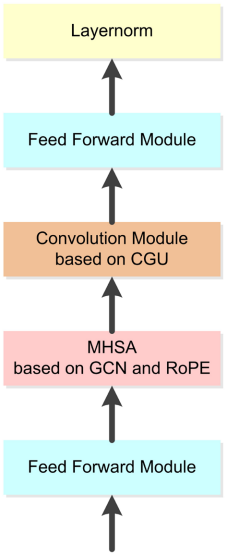

Figure 4. The structure of conformer block
图 4. Conformer Block 的结构图

3. 实验结果与分析

3.1. 数据集

为了评估所提出方法的有效性，本文使用 VoiceBank-DEMAND 任务[33]，该任务包含 Voice Bank 语音数据集[34]和来自 DEMAND 数据库[35]的噪声源。对于 Voicebank 数据集，训练集包括来自 28 位说话者的 11,572 个话语，测试集包括来自 2 位说话者的 824 个话语。训练集中的话语在四种不同的信噪比(SNR)下被来自 DEMAND 数据库的十种噪声污染：0、5、10 和 15 dB。测试集在四种信噪比下被五种 DEMAND 噪声污染：2.5、7.5、12.5 和 17.5 dB。特别地，从训练集中留出了大约 770 个话语用于作为验证集使用。

为了进一步探究所提出方法的泛化能力，本研究构建了另一个测试集。该测试集使用与原始测试集相同的无噪声语音片段，同时使用来自 NoiseX-92 数据集[36]的不同类型的噪声对其进行污染，以模拟更严重的噪声场景。使用的噪声是来自 NoiseX-92 数据集的三种噪声 babble、pink 及 f16。由此产生的噪声片段信噪比(SNR)分别为-5、0、5 和 10dB。值得注意的是，所有用于实验的波形均以 16 kHz 的采样率重新采样。

3.2. 实验装置

本研究提出的模型在 VoiceBank-DEMAND 任务上训练 30 个 epoch，同时使用 adam 进行优化，其步长设置为 3e−4，梯度动量和分母动量分别设置为 0.9 和 0.999。批大小设置为 24。为了更好地说明提出模型的性能，实验中使用了与 denoiser 方法一样的损失函数。它就是波形上的平均绝对误差以及对频谱幅度采用的多分辨率短时傅里叶变换(STFT)损失[37]。

验证实验采用语音质量感知评估(PESQ)、短时客观可懂度(STOI)和分段信噪比(SSNR)指标对增强效果进行评价。此外，还计算了语音失真预测值(CSIG)、背景噪声侵入度预测值(CBAK)及整体语音质量预测值(COVL)三个指标与其他算法在 VoiceBank-DEMAND 任务上进行实验方法的对比。

3.3. 实验结果

3.3.1. 与其他方法的比较

为了全面评价提出算法的增强性能，本实验采用了 denoiser 算法及它的变种 HDDEMUCS 进行实验对比，这些对比的网络模型均使用 VoiceBank-DEMAND 数据集进行实验验证。为直观地比较不同时域增强模型的效果，在分析各项评价指标后，还分析了噪声条件下增强语音的语谱图。表 1 描述了多种增强模型的对比实验结果。提出的模型的增强语音的各项评价指标得分除了 COVL 之外都要优于其他模型。与时域的 denoiser 方法相比，提出的模型在 PESQ 指标上提升了 0.0554，在 STOI 指标上提升了 0.0020，在 SSNR 指标上提升了 0.4389。同时，提出的模型的参数量相对更少一些。综合表中各项量化评估结果可见，本文所提出的算法在绝大多数核心性能指标上均优于当前主流时域模型。这一实验结果充分表明，该算法能够深度挖掘并有效利用序列数据中的时域关联信息，通过构建精细化的特征建模机制，实现了更为卓越的时域增强效果。

图 5 给出了带噪语音、干净语音和不同算法增强语音的语谱图。信号选自 VoiceBank-DEMAND 测试数据集中标签为 232 号说话人的第 5 条语音信号。从图 5 中可以看出，denoiser 算法在对语音段的增强没有提出的算法有效，依然保留有较多的噪声。对比图 5 黑色方框中的背景噪声及语谱分量可以发现，在低信噪比条件下，提出的算法对背景噪声的抑制更加充分，并且对语音谐波成分的保持更好。以上对比说明，提出的算法综合利用了 U-Net 架构的多尺度特征融合能力，以及卷积与注意力跳跃连接的特征补偿建模能力，对语音信号的建模更加准确。

Table 1. Experimental results of different methods on VoiceBank-DEMAND
表 1. 不同方法在 VoiceBank-DEMAND 数据库上的实验结果

方法名称	参数量(M)	PESQ	STOI	SSNR	CSIG	CBAK	COVL
Noisy	—	2.1163	0.9210	1.6758	3.3610	2.4429	2.6398
denoiser	32.01	2.7585	0.9314	9.6693	3.5677	3.1876	**2.9597**
HDDEMUCS	39.99	2.7750	0.9195	9.1263	3.4019	3.0699	2.8046
Proposed	29.99	**2.8139**	**0.9334**	**10.1082**	**3.5763**	**3.2051**	2.9553

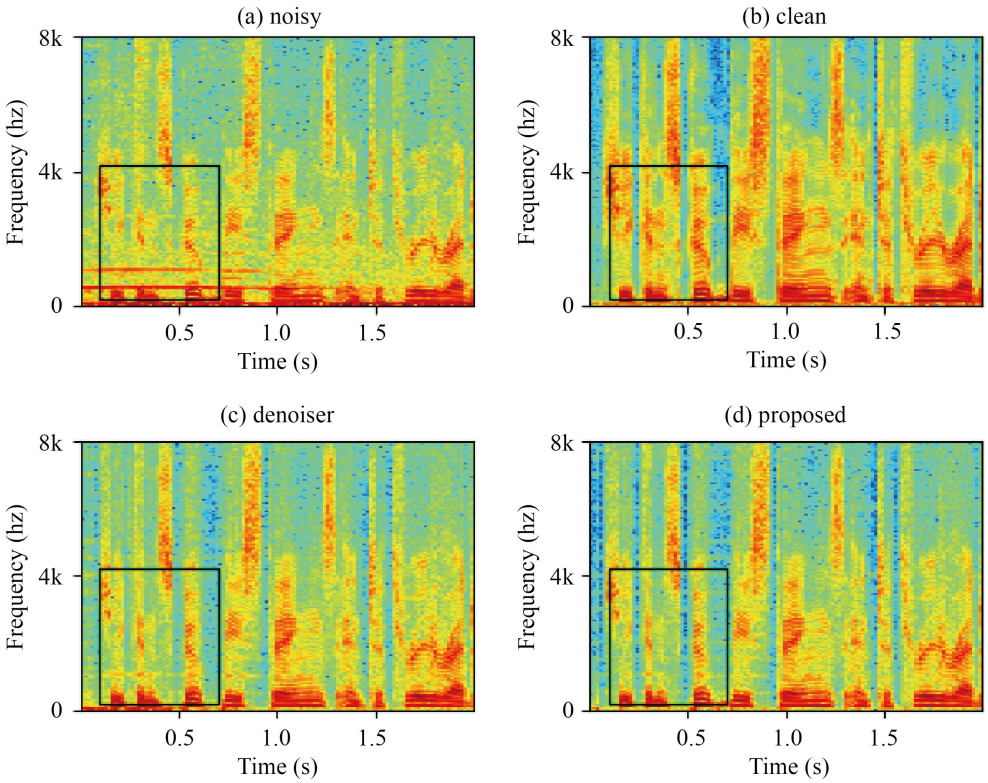

Figure 5. Spectrogram examples of different speech enhancement methods
图 5. 不同的语音增强方法的语谱图

Table 2. Experimental results of different methods under varying noise levels and signal-to-noise ratios
表 2. 不同方法在不同噪声不同信噪比下的的实验结果

	噪声类型	平稳噪声		非平稳噪声
	信噪比(−5 dB)	pink	f16	babble
PESQ	denoiser	1.3322	1.2371	1.1809
	HDDEMUCS	1.2964	1.2731	1.2378
	Proposed	**1.3331**	**1.3047**	**1.1948**
STOI	denoiser	0.6342	0.6094	0.6383
	HDDEMUCS	0.5756	0.5631	0.6343
	Proposed	**0.6920**	**0.6613**	**0.6682**

续表

SSNR	denoiser	1.4432	1.0006	0.9226
	HDDEMUCS	−0.4303	−0.6950	0.3231
	Proposed	**1.7249**	**1.4677**	**1.0136**
	信噪比(10 dB)	pink	f16	babble
PESQ	denoiser	2.1447	2.0781	**2.5523**
	HDDEMUCS	2.0376	2.1360	2.4276
	Proposed	**2.2888**	**2.3102**	2.5474
STOI	denoiser	**0.8883**	0.8704	0.8982
	HDDEMUCS	0.8737	0.8684	0.8887
	Proposed	0.8796	**0.8778**	**0.9009**
SSNR	denoiser	7.3849	7.2366	8.0792
	HDDEMUCS	6.9161	6.8693	7.9550
	Proposed	**7.6355**	**7.9934**	**8.6969**

表 2 列出了 2 种平稳噪声和 1 种非平稳噪声的测试语料的经过 3 中算法增强后的指标。观察表 2 可以看到，提出的算法相比于 denoiser 方法，对噪声的抑制程度更加明显。在低信噪比条件下，提出算法的所有指标都优于其他算法。在高信噪比条件下，提出算法的 SSNR 指标都优于其他算法。对于非平稳噪声而言，在低信噪比的情况下，提出的算法有更好的噪声抑制能力；在高信噪比的情况下，提出的算法只有 PESQ 指标略为偏低。

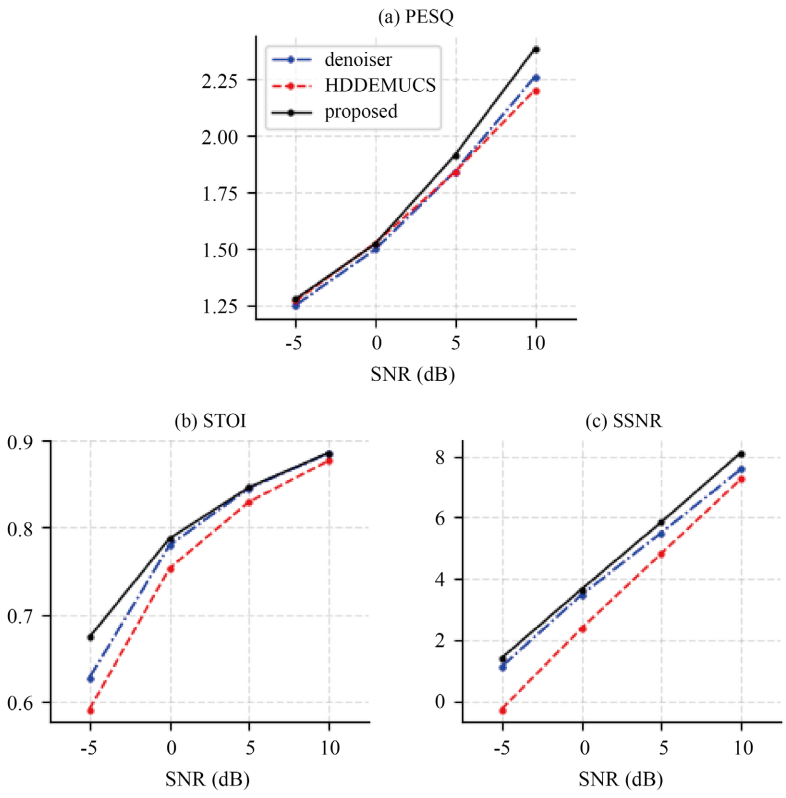

Figure 6. Indicator curves of three algorithms under different signal-to-noise ratios
图 6. 不同信噪比条件下三种算法的指标曲线

本实验还进一步对比分析了 denoiser、HDDEMUCS 以及提出的算法在不同信噪比条件下的增强效果，实验结果如图 6 所示。从该图中可以看出，PESQ 在高信噪比条件下，提出的算法有明显提升；而 STOI 在低信噪比条件下，提出的算法有明显提升。对于 SSNR 而言，提出的算法在各个信噪比条件下均有所提升。

3.3.2. 消融实验

表 3 中描述了不同模型的消融实验结果。第一个算法没有应用基于 CGU 的卷积模块以及基于 GCN 和 RoPE 的多头注意力机制，而是应用含门控线性单元(GLU)的卷积模块以及使用相对位置编码的多头注意力机制。第二个算法仅应用了基于 CGU 的卷积模块，它贡献了 PESQ 和 STOI 的提升。第三个算法仅利用了基于 GCN 和 RoPE 的多头注意力机制，它贡献了 PESQ 和 SSNR 的提升。最后，第四个算法同时使用了基于 CGU 的卷积模块以及基于 GCN 和 RoPE 的多头注意力机制，即提出的方法，它带来了全面的性能提升。

Table 3. Ablation experiment results of different models
表 3. 不同模型的消融实验结果

方法	PESQ	STOI	SSNR
raw conv + raw attention	2.7686	0.9307	9.9934
conv skip + raw attention	2.7695	0.9317	9.7934
raw conv + attention skip	2.7899	0.9305	10.0043
conv skip + attention skip (Proposed)	**2.8139**	**0.9334**	**10.1082**

4. 结语

本研究中提出了一种专为单通道语音增强设计的新型框架。该方法在 U-Net 网络中应用了基于卷积模块与注意力机制的跳跃连接，它借鉴了卷积门控单元、图卷积网络和旋转位置嵌入等技术。卷积门控单元专注于局部特征的学习以及重要信息的筛选。同时，图卷积网络捕捉了复杂的头部相关性。为了进一步提升性能，本算法还引入了一种旋转位置嵌入策略，该策略在多头自注意力模块中明确编码了相对位置信息。为了验证提出的方法的有效性，本研究使用了 VoiceBank-DEMAND 数据集进行实验验证，并采用了多种评估指标(PESQ、STOI 及 SSNR)进行评估。实验结果清晰地展示了我们的方法的优越性。未来，研究目标是以不同的方式将位置信息编码到输入表示中，以帮助模型更好地理解序列中元素的位置关系。此外，还将克服算法参数数量庞大的局限性。

参考文献

[1] Berouti, M., Schwartz, R. and Makhoul, J. (1979) Enhancement of Speech Corrupted by Acoustic Noise. *Proceedings of the* 1979 *IEEE International Conference on Acoustics*, *Speech*, *and Signal Processing* (*ICASSP*), Washington, 2-4 April 1979, 208-211.

[2] Ephraim, Y. (1992) Statistical-Model-Based Speech Enhancement Systems. *Proceedings of the IEEE*, **80**, 1526-1555. https://doi.org/10.1109/5.168664

[3] Lim, J. and Oppenheim, A. (1978) All-Pole Modeling of Degraded Speech. *IEEE Transactions on Acoustics*, *Speech and Signal Processing*, **26**, 197-210.

[4] Dendrinos, M., Bakamidis, S. and Carayannis, G. (1991) Speech Enhancement from Noise: A Regenerative Approach. *Speech Communication*, **10**, 45-57. https://doi.org/10.1016/0167-6393(91)90027-q

[5] Ephraim, Y. and Van Trees, H.L. (1995) A Signal Subspace Approach for Speech Enhancement. *IEEE Transactions on Speech and Audio Processing*, **3**, 251-266. https://doi.org/10.1109/89.397090

[6] Pascual, S., Bonafonte, A. and Serrà, J. (2017) SEGAN: Speech Enhancement Generative Adversarial Network. *Interspeech* 2017, Stockholm, 20-24 August 2017, 3642-3646. https://doi.org/10.21437/interspeech.2017-1428

[7] Cao, R., Abdulatif, S. and Yang, B. (2022) CMGAN: Conformer-Based Metric GAN for Speech Enhancement. *Interspeech* 2022, Incheon, 18-22 September 2022, 936-940. https://doi.org/10.21437/interspeech.2022-517

[8] Kim, M., Song, H., Cheong, S. and Shin, J.W. (2022) iDeepMMSE: An Improved Deep Learning Approach to MMSE Speech and Noise Power Spectrum Estimation for Speech Enhancement. *Interspeech* 2022, Incheon, 18-22 September 2022, 181-185. https://doi.org/10.21437/interspeech.2022-964

[9] Hwang, S., Park, S. and Park, Y. (2022) Monoaural Speech Enhancement Using a Nested U-Net with Two-Level Skip Connections. *Interspeech* 2022, Incheon, 18-22 September 2022, 191-195. https://doi.org/10.21437/interspeech.2022-10025

[10] Fu, Y., Liu, Y., Li, J., Luo, D., Lv, S., Jv, Y., *et al*. (2022) Uformer: A UNet Based Dilated Complex & Real Dual-Path Conformer Network for Simultaneous Speech Enhancement and Dereverberation. *ICASSP* 2022—2022 *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*), Singapore, 23-27 May 2022, 7417-7421. https://doi.org/10.1109/icassp43922.2022.9746020

[11] Wang, H. and Tian, B. (2025) ZipEnhancer: Dual-Path Down-Up Sampling-Based Zipformer for Monaural Speech Enhancement. *ICASSP* 2025—2025 *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*), Hyderabad, 6-11 April 2025, 1-5. https://doi.org/10.1109/icassp49660.2025.10888703

[12] Lee, S., Cheong, S., Han, S. and Shin, J.W. (2025) FlowSE: Flow Matching-Based Speech Enhancement. *ICASSP* 2025—2025 *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*), Hyderabad, 6-11 April 2025, 1-5. https://doi.org/10.1109/icassp49660.2025.10888274

[13] Gulati, A., Qin, J., Chiu, C., Parmar, N., Zhang, Y., Yu, J., *et al*. (2020) Conformer: Convolution-Augmented Transformer for Speech Recognition. *Interspeech* 2020, Shanghai, 25-29 October 2020, 5036-5040. https://doi.org/10.21437/interspeech.2020-3015

[14] Chen, Z., Yoshioka, T., Lu, L., Zhou, T., Meng, Z., Luo, Y., *et al*. (2020) Continuous Speech Separation: Dataset and Analysis. *ICASSP* 2020—2020 *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*), Barcelona, 4-8 May 2020, 7284-7288. https://doi.org/10.1109/icassp40776.2020.9053426

[15] 胡从刚, 申艺翔, 孙永奇, 等. 基于 Conformer 的端到端语音识别方法[J]. 计算机应用研究, 2024, 41(7): 2018-2024.

[16] Koizumi, Y., Karita, S., Wisdom, S., Erdogan, H., Hershey, J.R., Jones, L., *et al*. (2021) DF-Conformer: Integrated Architecture of Conv-Tasnet and Conformer Using Linear Complexity Self-Attention for Speech Enhancement. 2021 *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), New Paltz, 17-20 October 2021, 161-165. https://doi.org/10.1109/waspaa52581.2021.9632794

[17] Abdulatif, S., Armanious, K., Guirguis, K., Sajeev, J.T. and Yang, B. (2021) AeGAN: Time-Frequency Speech Denoising via Generative Adversarial Networks. 2020 28*th European Signal Processing Conference* (*EUSIPCO*), Amsterdam, 18-21 January 2021, 451-455. https://doi.org/10.23919/eusipco47968.2020.9287606

[18] Abdulatif, S., Armanious, K., Sajeev, J.T., Guirguis, K. and Yang, B. (2021) Investigating Cross-Domain Losses for Speech Enhancement. 2021 29*th European Signal Processing Conference* (*EUSIPCO*), Dublin, 23-27 August 2021, 411-415. https://doi.org/10.23919/eusipco54536.2021.9616267

[19] Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., *et al*. (2020) DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement. *Interspeech* 2020, Shanghai, 25-29 October 2020, 3885-3889. https://doi.org/10.21437/interspeech.2020-2537

[20] Défossez, A., Synnaeve, G. and Adi, Y. (2020) Real Time Speech Enhancement in the Waveform Domain. *Interspeech* 2020, Shanghai, 25-29 October 2020, 3291-3295. https://doi.org/10.21437/interspeech.2020-2409

[21] Kim, D., Chung, S., Han, H., Ji, Y. and Kang, H. (2023) HD-DEMUCS: General Speech Restoration with Heterogeneous Decoders. *INTERSPEECH* 2023, Dublin, 20-24 August 2023, 4125-4129. https://doi.org/10.21437/interspeech.2023-1642

[22] Wang, K., He, B. and Zhu, W. (2021) TSTNN: Two-Stage Transformer Based Neural Network for Speech Enhancement in the Time Domain. *ICASSP* 2021—2021 *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*), Toronto, 6-11 June 2021, 7098-7102. https://doi.org/10.1109/icassp39728.2021.9413740

[23] Défossez, A., Berrada, L., Dumoulin, V., *et al*. (2020) Music Source Separation in the Waveform Domain. arXiv: 1911.13254.

[24] 武瑞沁, 陈雪勤, 俞杰, 王丽荣, 赵鹤鸣. 结合注意力机制的改进 U-Net 网络在端到端语音增强中的应用[J]. 声学学报, 2022, 47(2): 266-275.

[25] 范君怡, 杨吉斌, 张雄伟, 郑昌艳. U-net 网络中融合多头注意力机制的单通道语音增强[J]. 声学学报, 2022,

47(6): 703-716.

[26] Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W. and Liu, Y. (2024) Roformer: Enhanced Transformer with Rotary Position Embedding. *Neurocomputing*, **568**, Article ID: 127063. https://doi.org/10.1016/j.neucom.2023.127063

[27] Bronstein, M.M., Bruna, J., LeCun, Y., Szlam, A. and Vandergheynst, P. (2017) Geometric Deep Learning: Going Beyond Euclidean Data. *IEEE Signal Processing Magazine*, **34**, 18-42. https://doi.org/10.1109/msp.2017.2693418

[28] Sadasivan, J., Seelamantula, C.S. and Muraka, N.R. (2020) Speech Enhancement Using a Risk Estimation Approach. *Speech Communication*, **116**, 12-29. https://doi.org/10.1016/j.specom.2019.11.001

[29] Cheng, J., Liang, R., Liang, Z., *et al*. (2023) A Deep Adaptation Network for Speech Enhancement: Combining a Relativistic Discriminator with Multi-Kernel Maximum Mean Discrepancy. *IEEE*/*ACM Transactions on Audio*, *Speech*, *and Language Processing*, **29**, 41-53.

[30] Hsieh, T., Wang, H., Lu, X. and Tsao, Y. (2020) WaveCRN: An Efficient Convolutional Recurrent Neural Network for End-To-End Speech Enhancement. *IEEE Signal Processing Letters*, **27**, 2149-2153. https://doi.org/10.1109/lsp.2020.3040693

[31] Yu, Z., Yu, L., Zheng, W. and Wang, S. (2023) EIU-Net: Enhanced Feature Extraction and Improved Skip Connections in U-Net for Skin Lesion Segmentation. *Computers in Biology and Medicine*, **162**, Article ID: 107081. https://doi.org/10.1016/j.compbiomed.2023.107081

[32] Kipf, T.N. and Welling, M. (2017) Semi-Supervised Classification with Graph Convolutional Networks. arXiv: 1609.02907.

[33] Valentini-Botinhao, C., Wang, X., Takaki, S. and Yamagishi, J. (2016) Investigating RNN-Based Speech Enhancement Methods for Noise-Robust Text-To-Speech. 9*th ISCA Workshop on Speech Synthesis Workshop* (*SSW* 9), Sunnyvale, 13-15 September 2016, 146-152. https://doi.org/10.21437/ssw.2016-24

[34] Veaux, C., Yamagishi, J. and King, S. (2013) The Voice Bank Corpus: Design, Collection and Data Analysis of a Large Regional Accent Speech Database. 2013 *International Conference Oriental COCOSDA Held Jointly with* 2013 *Conference on Asian Spoken Language Research and Evaluation* (*O-COCOSDA*/*CASLRE*), Gurgaon, 25-27 November 2013, 1-4. https://doi.org/10.1109/icsda.2013.6709856

[35] Thiemann, J., Ito, N. and Vincent, E. (2013) Demand: A Collection of Multi-Channel Recordings of Acoustic Noise in Diverse Environments. *Proceedings of Meetings on Acoustics*, Paris, 2-7 June 2013, 1-8.

[36] Varga, A. and Steeneken, H.J.M. (1993) Assessment for Automatic Speech Recognition: II. NOISEX-92: A Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems. *Speech Communication*, **12**, 247-251. https://doi.org/10.1016/0167-6393(93)90095-3

[37] Yamamoto, R., Song, E. and Kim, J. (2020) Parallel WaveGAN: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram. *ICASSP* 2020—2020 *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*), Barcelona, 4-8 May 2020, 6199-6203. https://doi.org/10.1109/icassp40776.2020.9053795