

Received 20 June 2024, accepted 8 July 2024, date of publication 16 July 2024, date of current version 24 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3429398

DMFNet:一种新型的自监督动态多聚焦语音降噪网络

杨承昊¹,陶毅¹,刘敬刘²,许肖梅¹

¹厦门大学应用海洋物理与工程系, 厦门, 361005, 中国

²北京化工大学信息科学与技术学院, 北京, 100029, 中国

本研究部分得到了国家自然科学基金(批准号: 41976178)的支持, 部分得到了福建省哲学社会科学规划基金(批准号: FJ2021 MJDZ003)资助的“美好生活与高质量发展研究之生态视角”重大项目的支持。

ABSTRACT 近年来, 神经网络的快速发展为语音去噪带来了显著提升。然而, 这些模型需要大量成对的噪声-干净语音数据进行监督训练, 限制了它们的广泛应用。尽管已有研究尝试使用仅含噪声的语音数据来训练去噪网络, 但现有的自监督方法往往存在连续性不足、降噪性能欠佳或高度依赖噪声建模的问题。在本工作中, 我们提出了一种高效的自监督动态多聚焦网络(DMFNet), 这是一种仅用噪声数据进行训练的语音去噪网络, 采用多尺度连接的编码器-解码器架构。具体而言, 我们设计了高效的频谱动态聚焦单元(SDFU), 使网络能够在学习特征时动态调整卷积核的形状, 从而精准聚焦于语音的频谱结构。此外, 我们引入了复杂注意模块(CAM), 该模块在跨空间结构设计中专注于特征的交互和提取。为进一步增强细微频谱细节的恢复, 我们提出了复杂多尺度特征融合单元(CMFFU)和复杂范围融合单元(CSFU), 以自适应地融合编码过程中不同阶段的特征。多个数据集上的广泛评估表明, 所提出的DMFNet在性能上显著优于其他最先进的方法。

INDEX TERMS 语音降噪, 音频子采样器, 自监督, 深度学习。

I. INTRODUCTION

音频去噪[1]的目标是从录制的语音中分离背景噪音, 以保持语音的感知质量和清晰度。这在音频通话、语音识别、设备通信和生物监测等应用中尤为重要。然而, 现实环境中的噪音不可避免地会影响目标信号, 常见的环境噪声包括人体噪声、空气噪声和水下噪声等。早期, 音频去噪主要采用传统信号处理方法, 如频谱减法[2]、维纳滤波[3]和小波去噪[4]。然而, 这些方法在处理非平稳或结构化噪声时效果有限, 尤其是针对如汽车喇叭声、儿童玩耍声和狗叫声等复杂噪声。

自20世纪80年代起, 神经网络就已应用于语音增强领域[5]。随着神经网络技术的不断发展, 语音增强的效果也取得了显著进步。这类模型通常在有监督的环境中训练, 通过从含噪音的输入中预测出干净的音频信号, 达到去噪的目的, 这种策略被称为噪声-干净训练(NCT)。然而, NCT方法存在一些局限性, 特别是在室外和自然环境中, 难以获得用于标记的干净样本。在实际应用中, NCT监督去噪算法常常面临成本高昂和数据稀缺等问题。

为应对这一挑战, 一些研究人员提出了噪声对噪声(N2N)策略[6], [7], 即在特定噪声条件下构建训练数据对, 仅使用噪声语音数据对来训练去噪神经网络。Wisdom等人[8]和Fujimura等人[9]进一步提出了噪声训练(NerNT)策略, 利用相同噪声条件下的训练

数据对进行网络训练。尽管这些方法依赖于噪声条件一致的训练数据集, 但在实际场景中找到噪声条件完全相同的训练数据对并不现实。为克服这一限制, Neighbor2Neighbor策略从单个噪声信号生成网络训练输入和目标[10]–[13]。该方法通过特定设计的子采样器生成训练数据对, 使去噪效果可与噪声-干净训练

(Noisy-Clean Training)相媲美, 尤其在泊松噪声或高斯噪声环境下表现良好。然而, 这种子采样策略会破坏声谱的连续性: 子采样使同一时间范围内的采样点数量减少一半, 从而影响频谱的自然结构和纹理, 最终降低去噪语音的质量。

当语音信号经过短时傅里叶变换处理时, 其周期性特征(即基频及谐波)会在频域中形成显著的峰值。这些峰值在频谱图中呈现为垂直列, 频谱图则是跨时间的频率分布的直观表示[36]。准确重构这些信息对于提升语音信号的质量和特征尤为关键。然而, 在纯噪声训练场景下, Neighbor2Neighbor策略的子采样方法会严重破坏频谱图的局部细节和连续性。因此, 能够捕捉频谱图的自然流动性, 并在不同尺度上聚合频谱图的实部和虚部特征显得非常重要。为此, 我们在本研究中提出了一种高效的自监督动态多聚焦网络(DMFNet), 用于在缺乏干净训练数据的情况下进行语音去噪。我们的模型基于U-Net架构[14], 包含编码器和解码器模块。我们特别引入了频谱动态聚焦单元(SDFU), 通

过动态调整卷积核的形状来捕捉频谱特征。即使由于子采样导致频谱图的连续性受损，这种方法依然能使模型聚焦于柱状声谱的基本结构。在编码器和解码器的分支中，复杂注意模块（CAM）促进特征提取和实部、虚部间的交互。此外，我们引入复杂多尺度特征融合单元（CMFFU），取代传统的跳跃连接，以在网络不同阶段实现灵活的特征融合。同时，复杂范围融合单元（CSFU）连接编码器和解码器的瓶颈层，进一步增强实部和虚部特征，从而提升了DMFNet的去噪性能。总结而言，本工作的主要贡献如下：

- 我们设计了SDFU单元，使其在特征学习过程中动态调整卷积核的形状，从而使模型能够聚焦于人声频谱的基本结构属性。在SDFU的基础上，进一步提出了复杂注意模块（CAM），用于增强和促进子采样频谱图实部和虚部特征提取的交互。
- 我们提出了复杂范围融合单元（CSFU），由复杂全局范围多层感知器（CGMLP）模块和复杂局部注意（CLA）模块组成。该单元对实部和虚部的局部信息进行编码，并建模长程依赖关系，从而实现高质量的频谱重建。
- 我们引入了复杂多尺度特征融合单元（CMFFU），旨在融合网络中不同尺度和深度的密集特征。此操作可确保模型获得丰富的特征信息，从而增强其重构高质量频谱图的能力。
- 我们设计了DMFNet网络，专用于语音去噪，仅需噪声语音样本。实验结果表明，DMFNet在多项指标上均达到了最先进的性能。

II. RELATED WORK AND IMPROVEMENT

有监督的语音去噪方法通常用神经网络表示为 f_θ ，其作用是将噪声语音 n 映射到估计的干净语音 $f_\theta(n)$ 。这类方法通过成对的干净语音 s 及其对应的噪声语音 $n = s + \mu$ 来训练，其中 μ 示噪声成分。有监督去噪被称为“噪声到纯净”（N2C），传统的N2C方法依赖纯净的训练音频目标，通常采用 L_2 损失函数来解决以下优化问题：

$$\operatorname{argmin}_\theta L_{2,n2c} = \operatorname{argmin}_\theta \mathbb{E} \left[\|f_\theta(n) - s\|_2^2 \right] \quad (1)$$

A. NOISE2NOISE

此外，神经网络还可以通过使用同一干净语音的不同噪声观测结果进行训练。N2N方法基于成对的噪声语音记录，即 $n_1 = s + \mu_1$ 和 $n_2 = s + \mu_2$ ，其中 μ_1 和 μ_2 是相互独立的噪声向量。这些噪声向量从零均值分布中采样，且与输入无关。具体而言，在训练阶段，N2N方法使用噪声输入 $n_1 = s + \mu_1 \sim P$ 和 噪声目标 $n_2 = s + \mu_2 \sim Q$ 作为训练数据对。

$$\begin{aligned} L_{2,n2n} &= \mathbb{E}_{(n_1, n_2)} \left\{ \|f_\theta(n_1) - n_2\|_2^2 \right\} \\ &= \mathbb{E}_{(n_1, n_2, \mu_2 \sim Q)} \left\{ \|f_\theta(n_1) - (s + \mu_2)\|_2^2 \right\} \\ &= L_{2,n2c} + \mathbb{E}_{\mu_2 \sim Q} \left\{ \mu_2^2 \right\} + \operatorname{Var}(\mu_2) \end{aligned} \quad (2)$$

在N2N条件下，由于 $\mathbb{E}_{\mu_2 \sim Q} \{\mu_2\} = 0$ ，可以得到 μ_2^2 的期望值等于 μ_2 的方差与其期望平方之和。利用该原

理，可以对公式(2)中的第三项进行展开。样本分布的方差 $\operatorname{Var}(\mu_2)$ 等于总体方差除以样本大小，因此随着噪声训练数据集规模的增大，公式(2)中的第二项和第三项逐渐趋近于零，最终N2N的 $L_{2,n2n}$ 损失值接近于N2C的 $L_{2,n2c}$ 损失值。

$$\lim_{|\text{TrainingDataSet}| \rightarrow \infty} L_{2,n2n} = L_{2,n2c} \quad (3)$$

从理论上讲，在无限大数据集的情况下，N2N训练的性能与N2C训练相当。然而，实际上，由于训练集规模有限，N2N的性能略低于N2C。在白噪声的情况下，N2C在多个指标上略微优于N2N，包括信噪比（SNR）、判断信噪比（SSNR）、噪声比（SSNR）、窄带PESQ分数（PESQ-NB）、宽带PESQ分数（PESQ-WB）和短期客观智能（STOI）[7]。然而，在UrbanSound8K数据集的某些类别中（如警报器声）[15]，N2N的表现超过了N2C，这得益于其出色的泛化能力，能够避免局部最优化[15], [16]。

B. NEIGHBOR2NEIGHBOR

尽管N2N的性能令人印象深刻，但其实际应用常常受到限制。这些限制主要源于在相同静态场景中获取成对的噪声语音训练数据所面临的挑战。例如，录音中的环境噪声可能会迅速变化，这使得数据的匹配变得困难。

为了克服这一限制，Huang等人[10]提出了Neighbor2Neighbor（Nbr2Nbr）方法，该方法通过使用随机邻域子采样器生成子采样训练数据对。这一方法可以在不依赖于完全匹配的噪声语音数据对的情况下进行训练，具有更高的灵活性。在Nbr2Nbr方法的基础上，Zhu等人[11]进一步开发了仅噪声训练（ONT）策略，将这一方法扩展到音频领域，从而为噪声去除提供了更加普适的解决方案。

具体来说，一对子采样器 $G = (g_1, g_2)$ 从单个噪声数据 n 生成一对噪声训练数据 $(g_1(n), g_2(n))$ 。与N2N不同的是，两个采样的噪声信号 $(g_1(n), g_2(n))$ 的真实值是不同的，即 $\varepsilon = \mathbb{E}_{(n,s)}(g_2(n)) - \mathbb{E}_{(n,s)}(g_1(n))$ 。通过引入 $g_1(n) = g_2(s) + \varepsilon + \eta$ ，其中 η 是噪声项，并假设 $\mathbb{E}(\eta) = 0$ ，我们可以得到：

$$\begin{aligned} &\mathbb{E}_{(n,s)} \|f_\theta(g_2(n)) - g_1(n)\|_2^2 \\ &= \mathbb{E}_{(n,s)} \|f_\theta(g_2(n)) - g_2(s) - \varepsilon - \eta\|_2^2 \\ &= L_{2,n2n} - \mathbb{E}_{(n,s)} [2\varepsilon \cdot (f_\theta(g_2(n)) - g_2(s))] \end{aligned} \quad (4)$$

考虑到理想的去噪器 f_θ^* 是用干净的数据训练并采用 ℓ_2 -损失的，它可以确保在给定噪声 n 的情况下，对于 $\ell \in \{1, 2\}$ ，满足 $f_\theta^*(n) = s$ 和 $f_\theta^*(g_\ell(n)) = g_\ell(s)$ 。因此，在最优网络 f_θ^* 情况下，以下结论成立：

$$\begin{aligned} &\mathbb{E}_{(n,s)} \{f_\theta^*(g_1(n)) - g_2(n) - (g_1(f_\theta^*(n)) - g_2(f_\theta^*(n)))\} \\ &= \mathbb{E}_{(n,s)} \{g_1(s) - g_2(n) - (g_1(s) - g_2(s))\} \\ &= \mathbb{E}_{(n,s)} \{g_2(s) - g_2(n)\} = 0 \end{aligned} \quad (5)$$

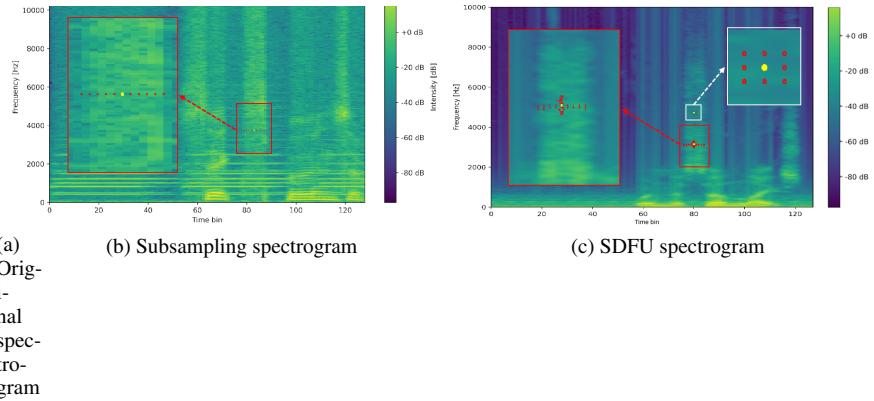


FIGURE 1. Changes in spectrograms after subsampling and visual representation of the proposed SDFU adaptive convolutional shape

C. IMPROVEMENT

Nbr2Nbr 方法通过引入子采样器部分解决了 N2N 方法的局限性。然而，这一子采样过程也带来了新的挑战，即信号的频谱图在子采样后缺乏结构连续性。子采样会破坏频谱图中结构和纹理的自然流动，导致信号的时间和频率信息被割裂，从而影响去噪音频信号的整体质量和真实感。这种结构连续性的缺失不仅影响去噪的效果，也削弱了模型对语音信号自然特征的保真度。

1) Horizontal direction (time axis)

在短时傅里叶变换 (STFT) 中，每个水平点对应一个时间帧。公式(5)中的 Hop_length 表示连续STFT 窗口之间的采样点数，决定了时间帧之间的间隔。对信号进行子采样后，相邻STFT 帧之间的时间间隔会比原始信号加倍，从而减少了表示相同时间跨度的帧数。这种子采样降低了时间分辨率，使得信号表达的信息量减少，导致频谱细节和连续性丢失，进而影响去噪信号的质量和保真度。

$$T = \frac{\text{Hop_length}}{\text{Sample_rate}}, \quad T' = \frac{\text{Hop_length}}{\text{Sample_rate}/2} = 2T \quad (6)$$

图1(a)展示了原始音频信号的频谱图，其中红色方框突出显示了沿特定时间轴的语音柱状频谱，捕捉了20个时间帧的语音频谱信息。而在图1(b)中，红色方框显示经过子采样后，相同的语音频谱信息被压缩到仅10个时间帧内。子采样明显降低了频谱图的时间分辨率，使得原始细节部分丢失，影响了频谱的连续性和对语音信息的捕捉能力。

2) Vertical direction (frequency axis)

在STFT 频谱图中，每个垂直点代表同一时间段内的不同频率成分，展示了特定时刻各频率的能量或功率分布。频谱图中这些频率点的间距决定了频率成分的分辨能力，且由FFT 窗口大小（参数 N_{FFT} ）控制。当子采样率减半时，频谱图中的频轴表示会被压缩，因为频率范围缩小。此时，原本分隔开的频点在垂直方向上更加接近，相邻的频率成分变得难以区分，影响频谱的清晰度。这是因为每个频点覆盖的频带范围变宽，从而降

低了频率分辨率。此外，超过子采样后奈奎斯特频率一半的频率成分可能因抗混叠滤波而被消除，或因混叠效应而出现在频谱的低频部分。这些变化对去噪后的信号质量和准确性造成影响，使得频谱特征的表达力降低。

$$\Delta f = \frac{\text{Sample_rate}}{N_{\text{FFT}}}, \quad \Delta f' = \frac{\text{Sample_rate}/2}{N_{\text{FFT}}} \quad (7)$$

因此，我们引入了SDFU 来解决子采样后频谱图信息丢失的问题。传统卷积核（如图1c 中白色方框所示）在水平和垂直方向上均无法有效捕捉子采样后频谱图的变化特征。图1c 中的红框展示了我们的方法，它通过自适应地调整卷积采样点的位置，使网络更高效地学习频谱图的细节信息。

在SDFU 中，黄色点代表卷积核的中心，而红色点则表示卷积的采样点。水平方向上，SDFU 显著增加了采样点的密度，并根据柱状声谱的结构自适应调整覆盖范围，从而准确捕捉沿时间轴的语音柱状结构。垂直方向上，SDFU 通过增加采样点密度，弥补了子采样后频谱压缩导致的信息缺失。这种自适应采样方式使得网络能够更准确地捕捉柱状声谱特征，从而改善频谱的自然结构和纹理流的呈现效果。

总之，SDFU 有助于增强网络的频谱特征学习能力，从而提升去噪音频信号的质量和保真度。更多关于SDFU 及其他网络增强模块的细节，详见第3 节。

III. SELF-SUPERVISED DYNAMIC MULTI-FOCUSING NETWORK

在本节中，我们将详细介绍所提出的DMFNet 的整体架构。首先，概述DMFNet 的整体结构，然后逐一介绍网络中的关键模块，最后说明用于自我监督DMFNet 训练的损失函数。

A. OVERVIEW

如图2 所示，DMFNet 接收噪声语音信号 S_{noise} 并通过子采样器生成一对噪声训练数据 $(S_{\text{noise1}}, S_{\text{noise2}})$ 。其中，子采样信号 S_{noise1} 经STFT处理，得到频域表示，作为训练网络的输入信号。网络输出去噪后的信号 $S_{\text{denoised1}}$ 与 S_{noise2} 计算损失，以更新网络权重。DMFNet 的架构基于U型对称分层结构，主要由编

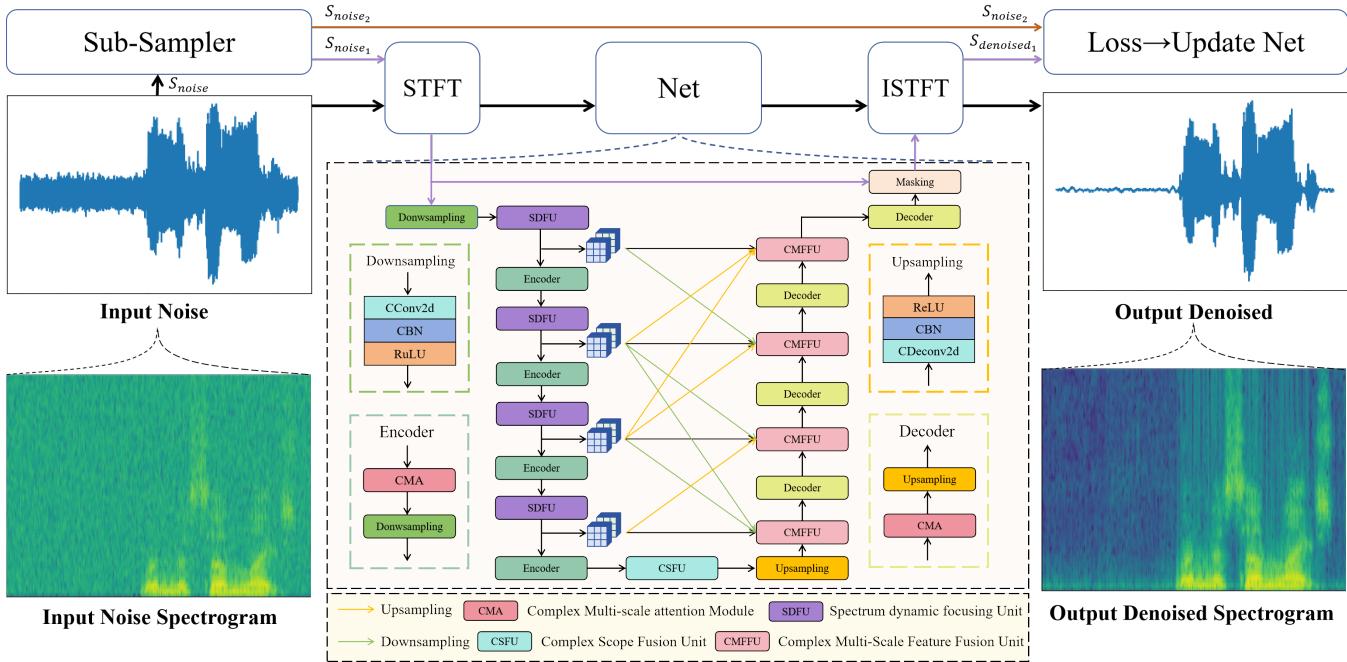


FIGURE 2. The complete structure of the proposed self-supervised DMFNet is a U-shaped network divided into two main phases: encoding and decoding, connected by CSFU. In the encoding phase, the network dynamically extracts local and global features at different scales. The decoding phase combines these features to reconstruct the spectral characteristics.

码和解码两个阶段组成。这两个阶段通过CSFU连接，以实现高效的特征提取和融合。编码器和解码器分别利用SDFU 和CMFFU 等模块提取频谱图的丰富特征。

1) Encoding Stage

在DMFNet 的编码阶段，网络包括四个连续的编码器，每个编码器负责提取不同频谱尺度的信息。在每个编码器前应用SDFU，该单元能够动态聚焦于输入语音的特定频谱区域（如人声频谱），帮助网络在缺乏结构连续性的条件下更准确地学习目标信号的频谱特征。每个编码器包含一个专门设计的复杂注意模块（CAM）和一个下采样块。下采样块由复数卷积层、复杂批量归一化层（CBN）、以及复杂ReLU 激活函数（CReLU）组成。复数卷积层使用复数卷积滤波器 $W = (A + iB)$ 进行计算。对于给定的复数向量 $h = c + id$ ，复数卷积通过两个独立的实数卷积操作来实现，其计算公式为： $W * h = (A * c - B * d) + i(B * c + A * d)$ 该架构中使用步长为2 的两个 3×3 实卷积核，在降低维度的同时进行复数卷积操作以提取特征。此外，CBN 和CReLU 进一步适应了复杂域中的特征处理，实验证明其在复数域上的性能优于许多现有方法 [17]。

2) Decoding stage

在DMFNet 的解码阶段，网络通过特征重建来恢复频谱图的局部和全局信息。为此，每个解码器前都应用了CMFFU，以帮助捕捉更丰富的频谱特征。每个解码器还包含CAM和一个上采样模块，用于优化特征的整合与重建。上采样模块采用复杂的转置卷积层，紧接着是CBN和CReLU。该转置卷积层由两个独立的实数反卷积实现，采用 3×3 核、步长为2，既能扩展特征维度，

又能在上采样过程中提取有效的特征信息。

B. SPECTRUM DYNAMIC FOCUSING UNIT (SDFU)

传统二维卷积核由于固定的几何形状，难以捕捉复杂频谱数据中的细节，尤其是在处理子采样音频信号时。因此，我们在此引入SDFU，如图3 所示。SDFU 通过融合传统二维卷积和新颖的柱状结构卷积（DSConv），可以更有效地检测人声频谱中的柱状局部特征。DSConv 通过有选择地关注柱状结构的局部几何属性，显著提升了复杂柱状结构的检测精度，并已在多个实验中证明其对复杂结构的辨别能力具有显著提升 [18], [19]。

图3 显示了SDFU 的输入和卷积核的配置，其中黄色区域显示了360 个红点，代表了经过学习后自适应选择的采样位置。这些点的可视化展示了SDFU 卷积核如何动态调整，以适应人声频谱中的柱状结构，确保卷积核与目标频谱区域的精确对齐。SDFU 的具体实现概述如下。

对于中心坐标为 (x_i, y_i) 的 3×3 卷积核 K_i ，标准二维卷积核 K 的定义如下：

$$K = \{(x_i - 1, y_i - 1), \dots, (x_i + 1, y_i + 1)\} \quad (8)$$

受 [20] 的启发，DSConv 在标准二维卷积核中引入了可学习的变形偏移 Δ ，从而使 K_i 沿着x 轴方向变成：

$$K_{i \pm c} = \begin{cases} (x_{i+c}, y_{i+c}) = (x_i + c, y_i + \sum_{k=i}^{i+c} \Delta y_k), \\ (x_{i-c}, y_{i-c}) = (x_i - c, y_i + \sum_{k=i-c}^{i} \Delta y_k). \end{cases} \quad (9)$$

同样 K_j 沿着y 轴方向变成：

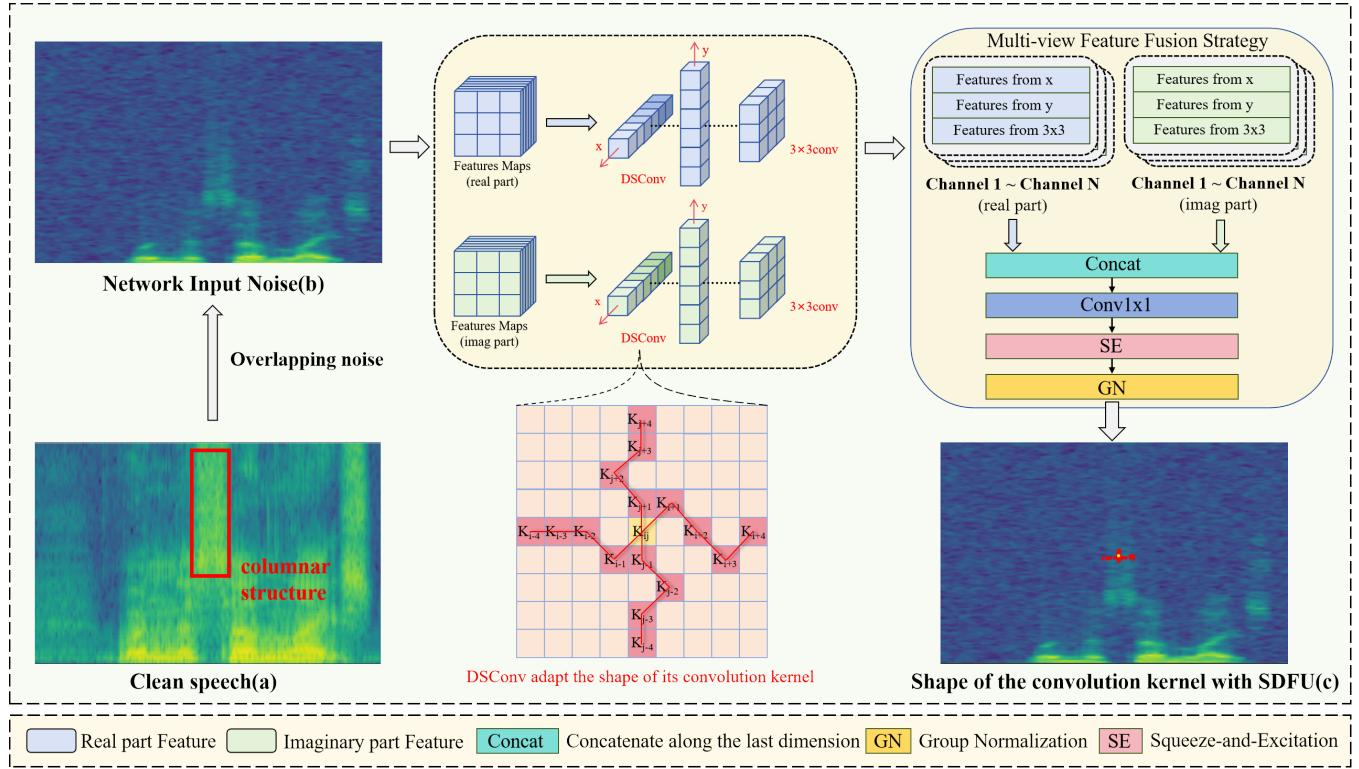


FIGURE 3. Figure (a) shows the original clean audio, and Figure (b) shows the audio with overlaid noise used as training data. Figure (c) shows the shape of the convolution kernel for a given sample point and illustrates the SDFU's ability to dynamically capture the human voice spectrum. The illustration features yellow dots representing the convolution kernels and 360 red dots marking the learning sample locations.

$$K_{j \pm c} = \begin{cases} (x_{j+c}, y_{j+c}) = (x_j + \sum_{k=j}^{j+c} \Delta x_k, y_j + c), \\ (x_{j-c}, y_{j-c}) = (x_j + \sum_{k=j-c}^j \Delta x_k, y_j - c), \end{cases} \quad (10)$$

其中，由于学习到的偏移量 Δx 和 Δy 通常不是整数，我们考虑采用采样双线性插值法，表示为：

$$K = \sum_{K'} B(K', K) \cdot K' \quad (11)$$

其中， K 表示方程(8)和(9)的分数位置， K' 表示整数空间中的所有位置， B 是双线性插值内核，它可以分成两个一维线性插值内核：

$$B(K, K') = b(K_x, K'_x) \cdot b(K_y, K'_y) \quad (12)$$

其中， b 表示一维线性插值核。传统的二维卷积能捕捉语音频谱的全局频谱特征。相比之下，DSConv可根据特征图在x轴和y轴方向上的形状对卷积核进行独特的调整。这种适应性提高了其准确检测语音频谱中局部柱状结构特征的能力，尤其是在不连续区域。通过DSConv和传统二维卷积提取的特征会沿着通道维度进行合并，然后进行 1×1 卷积进行初始融合，并通过组归一化处理进一步增强融合效果。随后，通过挤压-激发机制对每个通道进行加权，以提供定向关注。

假设STFT后网络的输入为 X_{in} ，则该过程可公式化描述为：

$$\hat{X} = \phi((DSC_x(X_{in}), DSC_y(X_{in}), Conv_{3 \times 3}(X_{in})) \quad (13)$$

其中 $\phi(\cdot) = Cat(SE(GN(Conv_{1 \times 1}(\cdot))))$ ， $DSC_x(\cdot)$ 和 $DSC_y(\cdot)$ 分别表示在x和y方向上的DSConv运算。 $Conv_{3 \times 3}(\cdot)$ 表示具有 3×3 内核的标准二维卷积。 $Conv_{1 \times 1}(\cdot)$ 表示使用 1×1 内核的标准二维卷积，其中输出通道的大小是输入通道的三分之一。GN表示组归一化，SE表示挤压-激发注意机制（Squeeze-and-Excitation mechanism）。[21]。

传统的二维卷积核往往难以捕捉复杂频谱数据中的细节特征，尤其是在子采样音频信号中。通过将传统的二维卷积与DSConv集成，SDFU有效地锁定并增强了对人声频谱中局部柱状结构的检测。DSConv引入了可学习的变形偏移，使卷积核能够动态适应柱状特征的特定几何形状。这种自适应方法可确保与目标频谱区域精确对齐，从而提高对复杂柱状结构的辨别能力。结合使用DSConv和传统的二维卷积，再加上组归一化和挤压-激发机制，有助于整合全局和局部频谱特征，从而提高语音频谱分析中特征提取的整体精度和效率。

C. COMPLEX ATTENTION MODULE(CAM)

CAM是DMFNet中的重要模块，用于局部和全局特征提取，并促进实部和虚部之间的信息交换。如图4所示，CAM主要包括复杂特征交互单元和跨空间学习单元，分别用于特征交互和特征提取。

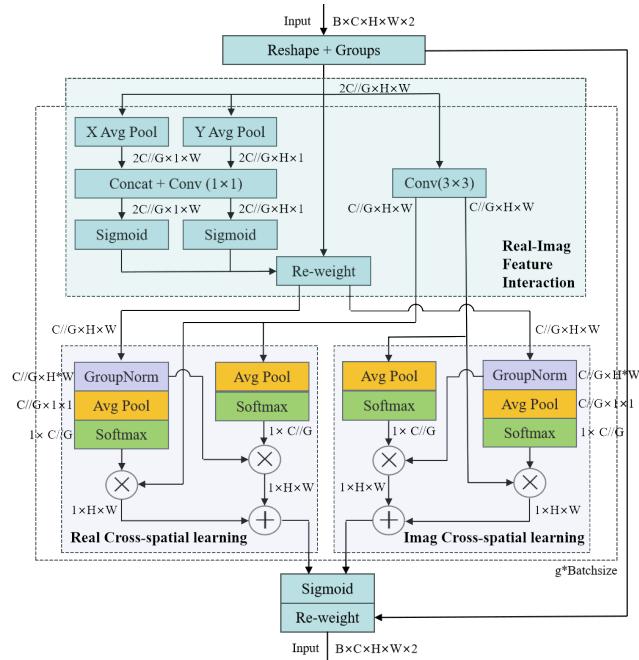


FIGURE 4. The architecture of CAM

1) Feature Grouping

对于任何给定的频谱输入 $J \in \mathbb{R}^{B \times C \times H \times W \times 2}$, 最后一维对应的是复数输入的实部和虚部。最初, CAM 沿着最后一个维度分割输入, 然后沿着通道维度将其连接起来。接下来, CAM 会沿着通道维度将输入划分为 $G \times \text{Batchsize}$ 个子特征, 从而加强对不同语义表征的学习。组风格的定义如下

$$I = [I_0, \dots, I_i, \dots, I_{G-1}], I_i \in \mathbb{R}^{2C//G \times H \times W} \quad (14)$$

在不失一般性的前提下, 我们让 $G \ll C$ 并假设学习到的注意力权重描述符将用于加强每个子特征中感兴趣区域的特征表示。

2) Complex Feature Interaction Unit

要对语音进行去噪, 首要挑战是有效利用噪声语音中的幅度和相位信息来重建清晰的信号。为了应对这一挑战, 我们引入了复杂特征交互单元 (CFIU)。该单元旨在最大限度地从语音信号的实分量和虚分量中提取有价值的信息并进行交互。如图4所示, CFIU 主要由三个并行的子结构组成: 两个以 1×1 分支为特征的共享分量实例和一个以 3×3 配置为特征的单一分支。

在 1×1 分支的共享分量中, 我们使用了一种类似于坐标注意 (CA) 模块中的一维特征编码向量的结构 [22]。该结构可捕捉实部和虚部横向和纵向的位置信息。全局平均池化技术可从输入张量的各个维度中提取这些向量。两个一维特征编码向量通过共享的 1×1 卷积层进行连接和处理, 以降低维度。这个卷积层是专门为有效捕捉局部跨通道交互而设计的。 1×1 卷积核的功能与通道卷积类似。卷积之后, 输出会在每条平行路径中通过一个非线性 sigmoid 函数。从这些并行路径中得出

的注意力权重会调整原始的中间特征图, 从而产生最终输出。

相反, 3×3 分支通过 3×3 卷积捕捉本地实数和虚数数据的跨通道交互, 从而增强了特征空间。这种方法允许 CMA 对通道间信息进行编码, 在实-虚串联后调整各种通道的重要性, 并保持通道内精确的空间结构信息。

3) Cross-spatial learning unit

通道-空间学习单元 (CSLU) 利用通道和空间维度之间的相互联系, 这一概念在现代计算机视觉任务中得到了广泛探索 [23] [24]。CSLU 利用通道和空间位置之间的短程和长程相关性, 这些相关性来自 CFIU 的 1×1 和 3×3 分支的输出, 分别记为 $R_{\text{real}1 \times 1}^{C//G \times H \times W}$ 和 $R_{\text{real}3 \times 3}^{C//G \times H \times W}$ 。这表明其在管理复杂空间结构方面具有先进的能力。

CSLU 将输入沿通道维度进行分割, 分别将其引导到实部和虚部的交叉空间学习单元中。 1×1 分支输出专注于实部和虚部之间的局部交互。随后, 我们使用二维全局平均池化 (GAP) 对 1×1 分支的输出 $R_{\text{real}1 \times 1}^{C//G \times H \times W}$ 和 3×3 分支的输出 $R_{\text{real}3 \times 3}^{C//G \times H \times W}$ 中的全局空间信息进行编码。二维全局池化操作定义如下:

$$\text{GAP}(I) = \frac{1}{H \times W} \sum_j \sum_i I(i, j) \quad (15)$$

这种设计编码了实部和虚部的全局信息, 并有效地模拟了远距离依赖性。随后, CSLU 将非线性函数应用于二维 GAP 输出, 使其与线性变换保持一致。CSLU 将并行处理输出 $R_{\text{real}1 \times 1}^{1 \times C//G}$ 和 $R_{\text{real}3 \times 3}^{C//G \times H \times W}$ 相结合, 并通过矩阵点积运算生成空间注意图, 收集不同尺度下的空间信息。同样, CSLU 生成另一组空间注意力图, 使用一个 S 型函数对这些权重进行聚合, 以创建最终的输出特征图。这个过程确保特征图准确地反映实部和虚部之间的跨空间依赖性, 从而保持空间信息的准确性。

通过整合跨空间信息和并行子结构的技术, CSLU 显著增强了网络处理复杂空间配置的能力, 特别是在捕捉像素级细节和真实与虚幻域的全局上下文方面。

D. COMPLEX SCOPE FUSION UNIT

如图5所示, CSFU 主要由两个并联单元组成。CGMLP 模块捕获并整合顶层特征 X_{top} 中实部和虚部的长程相关性。同时, CLA 模块应用于 X_{top} 时, 会聚合实部和虚部中的局部特征。这两个单元的特征图沿通道维度串联, 形成 CSFU 输出, 用于进一步特征提取。

用于在 X_{top} 和 CSFU 之间进行特征平滑的茎块包括两个 5×5 卷积, 分别有 90 个输出通道, 用于处理实部和虚部, 之后是批量归一化和激活层。该过程总结如下:

$$X_{\text{smooth}} = \text{Concat}(\phi(X_{\text{top_real}}), \phi(X_{\text{top_im}})) \quad (16)$$

其中, $\phi(\cdot) = \sigma(\text{BN}(\text{Conv}_{5 \times 5}(\cdot)))$ 。 $\text{Conv}_{5 \times 5}(\cdot)$ 表示 5×5 卷积运算, 步长为 1, 通道大小为 90, 用于实部和虚部。 $\text{BN}(\cdot)$ 是一个批量归一化层, 而 $\sigma(\cdot)$ 表示 ReLU 激活函数。最后, 沿最后一个通道维度将实部和虚部结果连接起来。

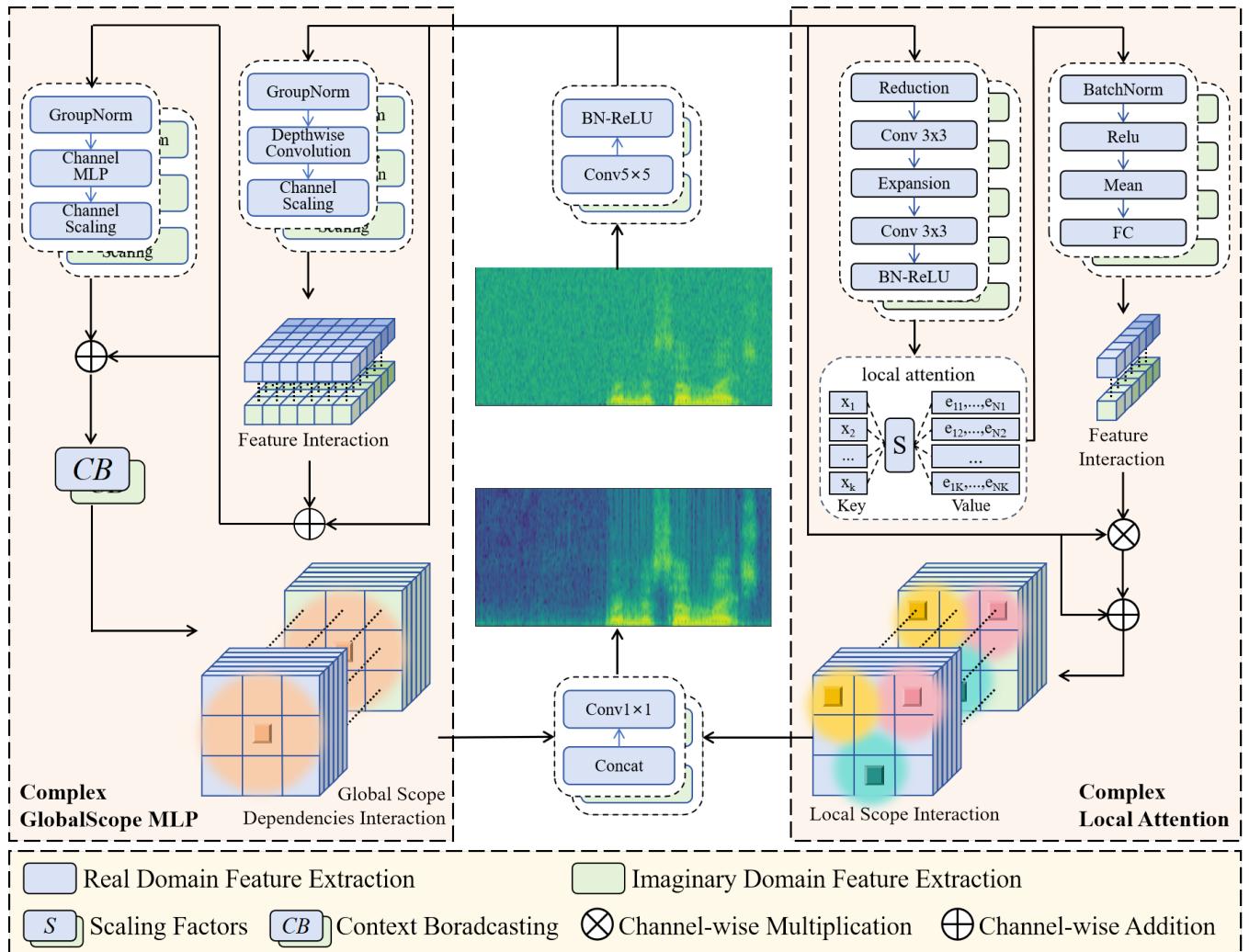


FIGURE 5. The architecture of the Complex Scope Fusion Unit. The CGMLP focuses on global interactions, while the CLA targets local corner interactions.

1) Complex GlobalScope Multi-Layer Perceptron

我们提出的CGMLP架构由三个主要部分组成：一个深度卷积残差模块、一个通道MLP残差模块和一个上下文广播（CB）模块。

深度卷积残差模块：在平滑了根模块后，输出特征 X_{smooth} 的实部和虚部均需进行深度卷积，然后进行分组归一化。与用于光谱特征提取的传统空间卷积不同，深度卷积在降低计算成本的同时增强了光谱特征的表示。随后，采用通道缩放操作来提高特征的泛化能力和鲁棒性。这些过程分别针对实部和虚部，其公式如下：

$$\tilde{X}_{dep} = CS(DConv(GN(X_{smooth}))) \quad (17)$$

其中， \tilde{X}_{dep} 表示深度卷积残差模块的输出。GN(\cdot) 表示组归一化，而DConv(\cdot) 表示深度卷积，其核大小为 1×1 。CS(\cdot) 表示通道缩放操作。

特征交互模块：在通道缩放操作之后，实部和虚部特征图沿通道维度串联。随后，使用 1×1 卷积来学习参数分布的压缩潜在空间表示，从而增强跨通道交互能力

[26], [27]。这些过程可以表述如下：

$$\tilde{X}_{1 \times 1} = Conv_{1 \times 1}(Cat(\tilde{X}_{dep_real}, \tilde{X}_{dep_im})) + X_{smooth} \quad (18)$$

其中，Cat(\cdot) 表示沿通道维度的连接，而Conv $_{1 \times 1}(\cdot)$ 表示 1×1 卷积。

通道MLP残差模块：首先，特征交互单元的特征（记为 $\tilde{X}_{1 \times 1}$ ）进行分组归一化，然后应用通道多层感知机（MLP）模型 [28]。通道MLP通过其多层结构有效地学习复杂的数据特征，同时降低计算复杂度 [29], [30]。随后，进行通道缩放操作，以改善特征泛化和鲁棒性。这些过程如下所示：

$$\tilde{X}_{mlp} = CS(MLP(GN(\tilde{X}_{1 \times 1}))) + \tilde{X}_{1 \times 1} \quad (19)$$

其中， \tilde{X}_{mlp} 表示通道MLP残差模块的输出。GN(\cdot) 表示组归一化，MLP(\cdot) 表示通道MLP。CS(\cdot) 表示通道缩放操作， $\tilde{X}_{1 \times 1}$ 表示特征交互模块的输出。

上下文广播模块：图6描述了位于CGMLP模块末端的CB。关于包含CB的整体CSFU架构，请参考图5。CSFU专注于涉及实部和虚部的密集交互，同时

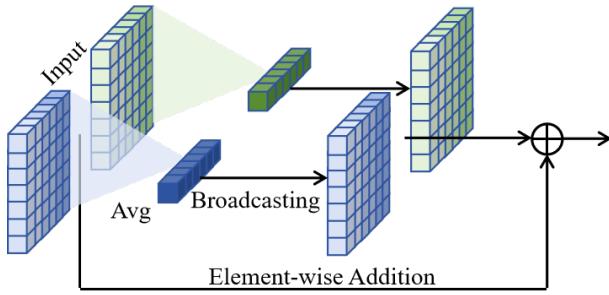


FIGURE 6. The architecture of CB

考虑长程和短程依赖性。然而，softmax函数的陡峭梯度使得密集注意力的学习变得复杂。相反，整合CB可以降低原始注意力图中的密度，从而增强CSFU的能力和泛化能力。具体来说，给定输入 $\tilde{X}_{\text{mlp}} \in \mathbb{R}^{C \times W \times H}$ ，CB对通道特征进行平均池化操作，如下所示：

$$\text{CB}(\tilde{X}_{\text{mlp}}) = \frac{\tilde{X}_{\text{mlp}} + \frac{1}{C} \sum_{j=1}^C X_{j,h,w}}{2} \quad (20)$$

其中， \tilde{X}_{mlp} 是复杂长程MLP模块的输出。 $\frac{1}{C} \sum_{j=1}^C$ 表示通道维度的平均池化操作。

2) The Complex Local Attention block

CLA模块作为编码器，内置字典机制。编码器由两个主要组件组成：局部注意单元和特征交互单元。

局部注意单元：如图5所示，输入特征首先通过一个组合模块进行处理，该模块包括 1×1 和 3×3 卷积。该模块旨在压缩特征、捕捉空间关系并扩展通道信息。然后，变换后的特征由CBR（卷积、批量归一化和ReLU）模块进行细化，该模块集成了 3×3 卷积、批量归一化和ReLU激活。这种整合对于增强特征描述和保持非线性至关重要。随后，编码后的特征被输入到局部注意单元。一组缩放因子用于对齐和映射相应的位置信息。与第k个编码字相关的所有特征的综合信息计算如下：

$$e_k = \sum_{i=1}^N \left[\frac{e^{-s \|a_{i,\text{real}} - b_{k,\text{real}}\|^2}}{\sum_{j=1}^K e^{-s \|a_{i,\text{real}} - b_{j,\text{real}}\|^2}} (a_{i,\text{real}} - b_{k,\text{real}}) + \frac{e^{-s \|a_{i,\text{imag}} - b_{k,\text{imag}}\|^2}}{\sum_{j=1}^K e^{-s \|a_{i,\text{imag}} - b_{j,\text{imag}}\|^2}} (a_{i,\text{imag}} - b_{k,\text{imag}}) \right] \quad (21)$$

其中， e_k 是第k个编码向量的更新。 s_k 表示一组缩放因子。 $a_{i,\text{real}}$ 和 $a_{i,\text{imag}}$ 分别对应于第*i*个输入向量的实部和虚部。 $b_{k,\text{real}}$ 和 $b_{k,\text{imag}}$ 分别对应于第*k*个编码向量的实部和虚部。我们使用特征融合单元来组合所有 e_k 并突出关键类别。该单元由一个BN层、一个ReLU激活函数、一个均值层和一个全连接层组成。在此基础上，所有*k*个编码字的完整信息计算如下：

$$e = \text{Conv}_{1 \times 1} \left(\sum_{k=1}^K \phi(e_k) \right) \quad (22)$$

其中， $\phi(\cdot)$ 表示BN层、ReLU激活函数和均值层，而 $\text{Conv}_{1 \times 1}$ 表示全连接层。

特征交互单元：在从特征融合单元获得输出后，我们引入了一个特征交互单元，用于实现局部特征的实部和虚部之间的交互。随后，我们对来自根模块的输入特征 X_{smooth} 和缩放因子系数 $\delta(\cdot)$ 进行逐通道乘法。最后，我们将来自根模块的特征 X_{smooth} 与局部关注特征进行通道级相加。上述过程总结如下：

$$\begin{aligned} \text{CLA}_{\text{out}} &= X_{\text{smooth}} \oplus \text{Cat}(X_{\text{smooth_real}} \otimes \delta(e_{\text{real}}), \\ &\quad X_{\text{smooth_im}} \otimes \delta(e_{\text{imag}})) \end{aligned} \quad (23)$$

其中， $\delta(\cdot)$ 表示sigmoid函数。 \otimes 表示通道级乘法。 \oplus 表示通道级加法。

CSFU旨在将声谱数据中的全局和局部特征无缝整合。它由两个主要模块组成：CGMLP和CLA模块。CGMLP通过一系列深度卷积、通道MLP和上下文广播模块，捕捉输入特征图实部和虚部的长程相关性，从而增强跨通道交互和特征表示。同时，CLA模块使用卷积和局部关注机制的组合方法，专注于局部特征聚合。这两个特征图沿通道维度串联起来，形成一个统一的输出，确保全面的特征提取。CSFU架构同时关注全局和局部交互，显著提高了模型捕捉复杂频谱特征的能力，从而提高了频谱数据分析的整体性能。

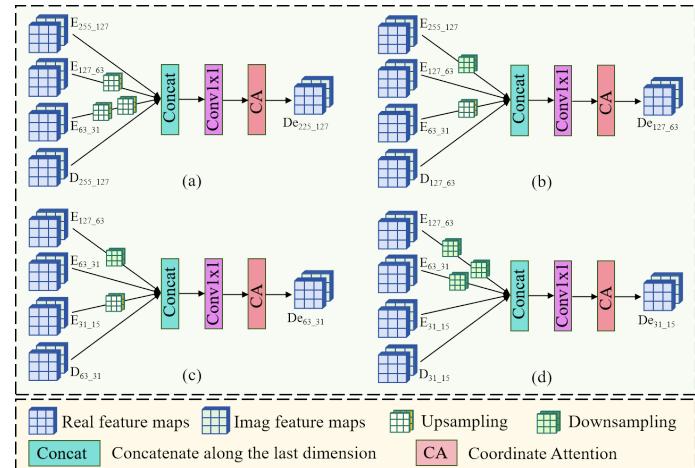


FIGURE 7. The architecture of CMFFU

E. COMPLEX MULTI-SCALE FEATURE FUSION UNIT

我们的首要目标是在解码过程中探索和利用编码阶段的实部和虚部特征。这些特征的大小不一，给整合带来了巨大挑战。例如，输入图像大小为 $512 \times 256 \times 2$ 时，解码阶段的特征图大小包括 $255 \times 127 \times 2$ 、 $127 \times 63 \times 2$ 、 $63 \times 31 \times 2$ 和 $31 \times 15 \times 2$ 。相反，解码阶段中的特征图大小将顺序颠倒为 $31 \times 15 \times 2$ 、 $63 \times 31 \times 2$ 、 $127 \times 63 \times 2$ 和 $255 \times 127 \times 2$ 。为了解决这一难题，我们引入了CMFFU，如图7所示。如图所示，解码阶段的每个特征图都与三个尺寸最相似的特征图合并。这个过程包括上采样和下采样，以调整特征图的大小，使其具有可比性，并利用不同的

TABLE 1. Speech denoising performance for different training strategies of artificial noise datasets.

Dataset	Network	SNR	SSNR	PESQ-NB	PESQ-WB	STOI
White Noise and 'Voice Bank + DEMAND'	NCT [7]	17.323 ± 3.488	4.047 ± 4.738	2.655 ± 0.428	1.891 ± 0.359	0.655 ± 0.017
	NNT [7]	16.937 ± 3.973	3.752 ± 4.918	2.597 ± 0.462	1.943 ± 0.375	0.645 ± 0.018
	SNA [37]	16.411 ± 1.837	3.283 ± 4.055	2.510 ± 0.285	1.805 ± 0.265	0.624 ± 0.179
	SDSD [38]	16.753 ± 1.672	3.808 ± 4.267	2.720 ± 0.271	1.813 ± 0.256	0.636 ± 0.183
	ONT [11]	17.563 ± 2.596	8.389 ± 2.961	2.690 ± 0.347	1.878 ± 0.293	0.833 ± 0.066
	ONT+rTSTM [11]	18.137 ± 2.122	9.077 ± 2.437	2.643 ± 0.317	2.003 ± 0.282	0.839 ± 0.067
	ONT+cTSTM [11]	18.209 ± 2.095	9.088 ± 2.222	2.811 ± 0.288	1.997 ± 0.276	0.847 ± 0.067
DMFNet(ours)		18.522 ± 2.159	9.413 ± 2.162	3.037 ± 0.320	2.203 ± 0.306	0.855 ± 0.073

采样层来处理实部和虚部。在标准化特征图大小后，我们将这些图沿通道维度连接起来进行整合。然后应用一个 1×1 卷积层来促进初始融合结果。最后，我们使用CA机制为每个通道分配一个特定通道的注意力权重。

CMFFU旨在充分利用从音频信号的STFT获得的频谱图的真实和虚部的多尺度特性。该单元通过利用编码阶段提取的关键信息来改善网络的特征传播和表示。CMFFU通过将特征图上采样和下采样到可比较的维度来应对整合不同大小特征的挑战。这些标准化的特征图沿着通道维度串联，然后进行 1×1 卷积，以进行初步融合和通道特定的注意力加权。这种全面的多尺度融合策略可确保对频谱数据的充分利用，这对于有效的自监督语音降噪至关重要。

F. LOSS FUNCTION

损失函数有三个项： ℓ_{base} 、 ℓ_{wSDR} [31]和 ℓ_{reg} [10]。我们将波形上的 ℓ_2 损失和STFT损失相结合，构建基本损失。 ℓ_{base} 的定义如下：

$$\ell_{base} = \frac{1}{N} \sum_{i=0}^{N-1} (n_i - \hat{n}_i)^2 + \frac{1}{TF} \sum_{t=0}^{T-1} \sum_{f=0}^{F-1} (|S_r(t,f)| - |\hat{S}_r(t,f)| - |\hat{S}_i(t,f)|) \quad (24)$$

其中， n_i 和 \hat{n}_i 分别表示第*i*对子语音和去噪子语音，而 N 是语音样本总数。 S 和 \hat{S} 分别表示子语音和去噪子语音的频谱。 r 和 i 是复变量的实部和虚部。 T 和 F 分别表示帧数和频率分帧数。

Choi等人提出的 ℓ_{wSDR} [31]是另一个术语，用于直接优化时域中定义的著名评估指标：

$$\ell_{wSDR} = -\alpha \frac{\langle n_1, \hat{n}_1 \rangle}{\|n_1\| \|\hat{n}_1\|} - (1-\alpha) \frac{\langle n_2 - n_1, n_2 - \hat{n}_1 \rangle}{\|n_2 - n_1\| \|n_2 - \hat{n}_1\|} \quad (25)$$

$$\alpha = \frac{\|n_1\|^2}{\|n_1\|^2 + \|n_2 - n_1\|^2} \quad (26)$$

其中， n_1 表示噪声子语音样本， \hat{n}_1 表示去噪后的子语音， n_2 表示另一个噪声子语音样本。 α 表示去噪后的子语音与另一个噪声子语音样本之间的能量比。

正如我们在第2.2节中描述的那样， ℓ_{reg} 是对损失函数的额外约束，其定义为：

$$\ell_{reg} = \|f_\theta(g_1(n)) - g_2(n) - (g_1(f_\theta(n)) - g_2(f_\theta(n)))\|_2^2 \quad (27)$$

其中，从含噪语音 n 中采样得到语音对 $g_1(n)$ 和 $g_2(n)$ ， f_θ 表示去噪网络。

IV. EXPERIMENTS AND RESULTS

A. DATASETS

在我们的实验中，我们创建了两个合成数据集，在静态说话者的清晰语音信号上叠加了各种类型的噪声，包括人工和真实世界的噪声。清晰的语音样本来自语音库数据集 [32]，其中包括来自28位说话者的训练样本和来自2位说话者的测试样本。

人工噪声数据集：第一个合成数据集包括高斯白噪声，噪声的音量经过调整，使得原始的信噪比在0到10（包括0和10）之间，从而产生一个盲去噪场景。然后使用PyDub [33]将噪声叠加在干净的音频上，PyDub截断或重复噪声，使其覆盖整个语音片段。

真实世界噪声数据集：第二个数据集包含来自UrbanSound8K (US8K) 的10个噪声类别，引用 [14]。在数据集混合过程中，我们还利用PyDub对各种噪声类别进行降噪，并使用UrbanSound8K通过10倍交叉验证来评估分类模型。

B. IMPLEMENTATION DETAILS

训练和测试语音信号的采样频率为48 kHz。我们使用PyTorch框架在NVIDIA GTX 4090 GPU上实现了我们的模型。我们使用Adam优化器优化了模型。学习率设置为 1×10^{-4} ，如果损失函数在特定周期内没有下降，则将其减半。

为了评估降噪质量，我们使用了以下指标：信噪比 (SNR)、信噪比 (SSNR)、PESQ-WB [34]、PESQ-NB [34]和STOI [35]。

C. QUANTITATIVE EVALUATION

表1和表2展示了我们的DMFNet在各项指标上的优势，结果表明，我们提出的DMFNet在人工噪声和真实噪声数据集上均取得了更好的结果。

1) Evaluation on artificial noise datasets.

如表1所示，结果表明，在人工噪声数据集上，我们的DMFNet在语音质量和信噪比方面优于其他模型，包括NCT的模型。在表1中，我

TABLE 2. Speech denoising performance for different training strategies of real-world noise datasets.

Dataset	Network	SNR	SSNR	PESQ-NB	PESQ-WB	STOI
'UrbanSound8K-0' (air conditioner) and 'Voice Bank + DEMAND'	NCT [7]	4.174 ± 3.608	-1.433 ± 3.124	1.980 ± 0.232	1.386 ± 0.165	0.578 ± 0.018
	NNT [7]	4.656 ± 5.612	-0.800 ± 3.687	2.440 ± 0.386	1.658 ± 0.298	0.641 ± 0.017
	NerNT [39]	4.318 ± 4.026	-1.294 ± 2.188	2.140 ± 0.332	1.160 ± 0.198	0.697 ± 0.018
	SNA [37]	1.324 ± 3.793	-5.216 ± 2.983	1.973 ± 0.449	1.250 ± 0.233	0.600 ± 0.180
	SDSD [38]	2.664 ± 2.447	-4.337 ± 2.456	2.022 ± 0.286	1.358 ± 0.237	0.552 ± 0.162
	ONT [11]	6.270 ± 3.711	1.185 ± 2.685	2.615 ± 0.488	1.776 ± 0.283	0.900 ± 0.09
	ONT+rTSTM [11]	6.231 ± 3.773	1.314 ± 2.704	2.143 ± 0.521	1.336 ± 0.300	0.809 ± 0.90
	ONT+cTSTM [11]	6.317 ± 3.813	1.414 ± 2.684	2.730 ± 0.485	1.891 ± 0.294	0.806 ± 0.09
	DMFNet(ours)	6.431 ± 3.652	1.446 ± 2.537	2.778 ± 0.465	1.943 ± 0.326	0.822 ± 0.084
'UrbanSound8K-1' (car horn) and 'Voice Bank + DEMAND'	NCT [7]	4.143 ± 3.899	-0.415 ± 3.664	1.924 ± 0.313	1.370 ± 0.208	0.562 ± 0.20
	NNT [7]	4.823 ± 6.166	0.324 ± 4.558	2.445 ± 0.481	1.770 ± 0.410	0.634 ± 0.19
	NerNT [39]	4.464 ± 3.858	-0.837 ± 3.714	2.121 ± 0.351	1.484 ± 0.256	0.651 ± 0.19
	SNA [37]	1.491 ± 3.786	-4.890 ± 2.983	1.765 ± 0.302	1.236 ± 0.158	0.563 ± 0.176
	SDSD [38]	2.165 ± 2.376	-4.407 ± 2.341	1.774 ± 0.268	1.275 ± 0.159	0.519 ± 0.161
	ONT [11]	6.244 ± 4.039	0.382 ± 4.029	2.650 ± 0.488	1.836 ± 0.324	0.759 ± 0.12
	ONT+rTSTM [11]	6.234 ± 4.051	0.505 ± 3.486	2.773 ± 0.518	1.861 ± 0.286	0.761 ± 0.12
	ONT+cTSTM [11]	6.339 ± 4.045	0.609 ± 3.160	2.954 ± 0.429	1.902 ± 0.376	0.850 ± 0.12
	DMFNet(ours)	6.422 ± 3.759	0.749 ± 2.982	2.856 ± 0.420	2.007 ± 0.336	0.852 ± 0.111
'UrbanSound8K-2' (children playing) and 'Voice Bank + DEMAND'	NCT [7]	3.830 ± 3.580	-1.403 ± 3.201	1.854 ± 0.235	1.332 ± 0.152	0.550 ± 0.17
	NNT [7]	4.348 ± 5.370	-0.636 ± 3.776	2.177 ± 0.378	1.512 ± 0.248	0.620 ± 0.17
	NerNT [39]	3.636 ± 3.392	-1.936 ± 3.103	1.812 ± 0.258	1.265 ± 0.134	0.572 ± 0.17
	SNA [37]	1.491 ± 3.786	-4.890 ± 2.983	1.765 ± 0.302	1.236 ± 0.158	0.563 ± 0.176
	SDSD [38]	2.165 ± 2.376	-4.407 ± 2.341	1.774 ± 0.268	1.275 ± 0.159	0.519 ± 0.161
	ONT [11]	6.559 ± 4.440	-0.343 ± 3.600	3.410 ± 0.504	1.963 ± 0.312	0.879 ± 0.10
	ONT+rTSTM [11]	6.546 ± 4.449	-0.287 ± 3.514	3.027 ± 0.502	1.806 ± 0.314	0.774 ± 0.10
	ONT+cTSTM [11]	6.442 ± 4.419	-0.302 ± 3.509	3.018 ± 0.503	1.867 ± 0.303	0.777 ± 0.10
	DMFNet(ours)	6.741 ± 4.353	-0.253 ± 3.442	2.882 ± 0.428	2.073 ± 0.303	0.754 ± 0.097
'UrbanSound8K-3' (dog bark) and 'Voice Bank + DEMAND'	NCT [7]	3.348 ± 3.457	-0.684 ± 3.767	1.773 ± 0.326	1.326 ± 0.190	0.520 ± 0.18
	NNT [7]	3.990 ± 5.451	-0.002 ± 5.084	2.147 ± 0.535	1.550 ± 0.372	0.593 ± 0.22
	NerNT [39]	3.537 ± 3.465	-1.336 ± 3.105	1.787 ± 0.260	1.249 ± 0.126	0.569 ± 0.17
	SNA [37]	1.752 ± 4.032	-4.445 ± 3.377	1.856 ± 0.408	1.264 ± 0.239	0.575 ± 0.183
	SDSD [38]	1.862 ± 2.885	-4.354 ± 2.592	1.822 ± 0.327	1.314 ± 0.225	0.527 ± 0.164
	ONT [11]	6.580 ± 6.687	3.982 ± 6.959	2.181 ± 0.712	1.599 ± 0.541	0.768 ± 0.17
	ONT+rTSTM [11]	6.592 ± 6.779	4.106 ± 7.386	2.193 ± 0.732	1.622 ± 0.591	0.772 ± 0.18
	ONT+cTSTM [11]	6.615 ± 6.886	4.199 ± 7.390	2.193 ± 0.735	1.626 ± 0.603	0.773 ± 0.17
	DMFNet(ours)	6.849 ± 6.454	4.313 ± 7.116	2.332 ± 0.320	1.783 ± 0.306	0.785 ± 0.153

TABLE 3. Ablation study

Network	SNR	SSNR	PESQ-NB	PESQ-WB	STOI
DCUnet [11]	17.663	8.604	2.751	1.936	0.834
DMFNet-1	18.235	9.099	2.990	2.138	0.849
DMFNet-2	18.417	9.077	3.006	2.159	0.839
DMFNet-3	18.466	9.272	3.014	2.143	0.847
DMFNet(ours)	18.522	9.413	3.037	2.203	0.855

们的DMFNet与其他方法相比取得了卓越的降噪效果，SNR为18.522 dB，SSNR为9.413 dB，PESQ-NB为3.037 dB，PESQ-WB为2.203 dB，STOI为0.855，彰显了其卓越的降噪性能。结果表明，我们的方法能够更有效地捕捉柱状人声频谱结构以及频谱图实部和虚部的独特特征。其中一个可能的原因是，在低信噪比下，DMFNet能够有效地将注意力集中在柱状人声频谱结构上。

2) Evaluation on real world noise datasets.

我们的DMFNet在真实噪声数据集上取得了最

佳结果。如表2所示，我们的评估集中在US8K中的四种真实加性噪声：空调、汽车喇叭、儿童游戏声和狗叫声。DMFNet在US8K类别0（空调）、1（汽车喇叭）和3（狗叫声）中几乎取得了最佳结果。ONT在US8K类别2中表现优于DMFNet，但差异并不明显。在US8K中，类别2主要包含儿童喊叫声，而DMFNet在去噪过程中可能保留了部分人声的频谱结构。结果表明，我们的DMFNet在合成和真实噪声数据集上均具有出色的去噪性能，特别是在非语音干扰的情况下。

D. VISUAL RESULTS

当使用STFT处理语音信号时，信号的周期性（由基频及其谐波决定）会在频域中产生明显的峰值，在频谱图中表现为竖条。为了证明DMFNet的有效性，图8显示了解码器处理前后的频谱。图8（a）清楚地表明，在解码之前，柱状声谱被严重的噪声干扰所淹没。特别是，在子采样之后，子信号的连续性丧失使得网络更难识别精确处理所需的人声特征。

图8（b）显示，DCUNet的第一个编码器输出有效

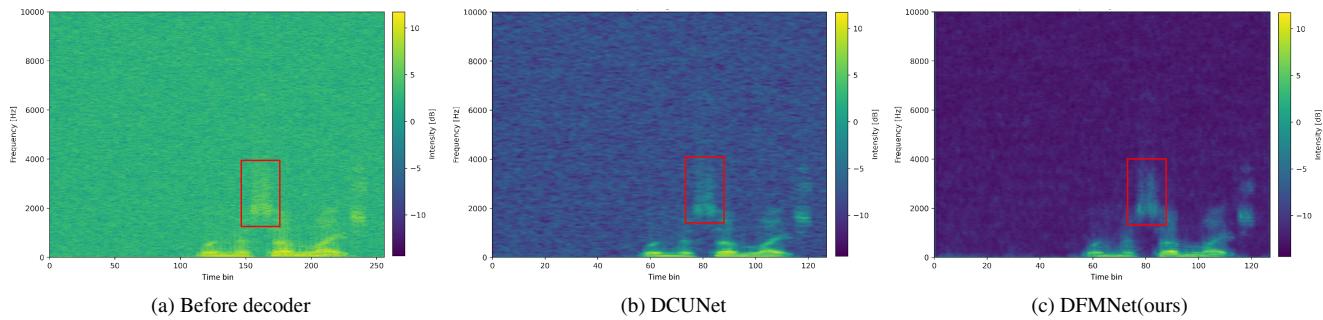


FIGURE 8. Fig. 8(a) displays the original noise-containing signal spectrum before encoding. Fig. 8(b) illustrates the output from the DCUNet encoder used in the ONT method, and Fig. 8(c) depicts the output from our proposed DMFNet encoder.

地提取了声谱的大部分特征。这表明DCUNet能够从噪声中分离出相关的频谱特征。然而，由于信噪比低，编码器输出仍然包含大量噪声干扰。

图8 (c) 展示了DMFNet的编码性能，这归功于SDFU能够专注于柱状声谱并准确捕捉目标频谱特征。SDFU在此过程中发挥了关键作用，因为它能够专注于与人类声音相对应的重要频谱柱，从而有效地过滤掉无关的噪声。这表明该网络能够在信噪比较低的环境中捕捉到详细特征。即使在具有挑战性的声学环境中，改进后的聚焦和特征捕捉功能也能更清晰、更准确地呈现人声频谱。与之前的方法相比，DMFNet的性能得到了提升，输出频谱更清晰、更独特，这表明它具有处理信噪比较低环境的卓越能力。

E. ABLATION STUDIES

本节介绍了消融研究，以展示所提出的DMFNet模型中不同模块的性能。我们选择ONT方法所使用的DCUNet [11]作为基准。这些模型被分为以下几类进行检验：(a) 应用CMA的DCUNet模型（用DMFNet-1表示）；(b) 应用CMA和CMFFU的DCUNet模型（用DMFNet-2表示）；(c) 应用CMA、CMFFU和CSFU的DCUNet模型（用DMFNet-3表示）。为了消除其他干扰，所有消融研究都是在叠加了白噪声的“语音库+DEMAND”上进行的。表3提供了SNR、SSNR、PESQ-NB、PESQ-WB和STOI值方面的总体结果。正如预期的那样，完整模型表现最佳。将CAM纳入模型显著提高了SNR、PESQ和STOI值。将SDFU整合到解码器阶段，可显著提高PESQ-NB和PESQ-WB值，这表明SDFU模块对于提高语音质量的重要性。

V. CONCLUSION

在这项工作中，我们介绍了一种新颖的自我监督DMFNet，它仅使用含噪语音数据来设计语音去噪。DMFNet采用多尺度连接编码器-解码器架构作为其骨干，并取得了非凡的成果。具体来说，我们设计了一种高效的SDFU，使网络能够有效地专注于柱状人声频谱结构，即语音去噪的目标频谱。此外，为了应对自监督子采样方法导致的连续性损失，我们引入了CMA模块，该模块同时关注局部频谱细节和全局频谱结构。此外，我们提出了CMFFU和CSFU，用于在编码器和解码器之间精细融合不同尺度和深度的特征。在人工和真实

数据集上的大量实验表明，DMFNet超越了其他现有方法，充分证明了其实用性和适用性。

ACKNOWLEDGMENT

这项工作部分得到了国家自然科学基金（批准号：41976178）的支持，部分得到了福建省哲学社会科学规划基金（批准号：FJ2021 MJDZ003）资助的“美好生活与高质量发展研究之生态视角”重大项目的支持。感谢高级工程师吴剑明在计算机平台方面的协助。

REFERENCES

- [1] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, 2007.
- [2] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.
- [3] J. S. Lim and A. V. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.
- [4] N. Priyadarshani, S. Marsland, I. Castro, and A. Punchihewa, “Birdsong denoising using wavelets,” *PloS One*, vol. 11, no. 1, e0146790, 2016.
- [5] S. Tamura and A. Waibel, “Noise reduction using connectionist models,” in *Proc. ICASSP*, vol. 1, pp. 553–556, 1988.
- [6] N. Alamdar, A. Azarang, and N. Kehtarnavaz, “Improving deep speech denoising by noisy2noisy signal mapping,” *Applied Acoustics*, vol. 172, 107631, 2021.
- [7] M. M. Kashyap, A. Tambwekar, K. Manohara, and S. Natarajan, “Speech denoising without clean training data: A noise2noise approach,” *arXiv preprint arXiv:2104.03838*, 2021.
- [8] S. Wisdom, E. Tzinis, H. Erdogan, R. Weiss, K. Wilson, and J. Hershey, “Unsupervised sound separation using mixture invariant training,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 3846–3857, 2020.
- [9] T. Fujimura, Y. Koizumi, K. Yatabe, and R. Miyazaki, “Noisy-target training: A training strategy for DNN-based speech enhancement without clean speech,” in *2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 436–440, 2021.
- [10] T. Huang, S. Li, X. Jia, H. Lu, and J. Liu, “Neighbor2neighbor: Self-supervised denoising from single noisy images,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14781–14790.
- [11] J. Wu, Q. Li, G. Yang, L. Li, L. Senhadji, and H. Shu, “Self-supervised speech denoising using only noisy audio signals,” *Speech Communication*, vol. 149, pp. 63–73, 2023, Elsevier.
- [12] J. Zhu, W. Cai, M. Zhang, and Y. Yang, “Self-supervised denoising model based on deep audio prior using single noisy marine mammal sound sample,” *Applied Intelligence*, vol. 53, no. 21, pp. 25697–25714, 2023, Springer.
- [13] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III*, vol. 18, Springer, 2015, pp. 234–241.

- [14] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM International Conference on Multimedia*, 2014, pp. 1041–1044.
- [15] N. Alamdar, A. Azarang, and N. Kehtarnavaz, "Improving deep speech denoising by noisy2noisy signal mapping," *Applied Acoustics*, vol. 172, pp. 107631, 2021, Elsevier.
- [16] M. Zhou, T. Liu, Y. Li, D. Lin, E. Zhou, and T. Zhao, "Toward understanding the importance of noise in training neural networks," in *International Conference on Machine Learning*, 2019, pp. 7594–7602, PMLR.
- [17] C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal, "Deep complex networks," *arXiv preprint arXiv:1705.09792*, 2017.
- [18] Q. Liu, Y. Liu, and D. Lin, "Revolutionizing Target Detection in Intelligent Traffic Systems: YOLOv8-SnakeVision," *Electronics*, vol. 12, no. 24, pp. 4970, 2023.
- [19] Y. Qi, Y. He, X. Qi, Y. Zhang, and G. Yang, "Dynamic snake convolution based on topological geometric constraints for columnar structure segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6070–6079, 2023.
- [20] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 764–773.
- [21] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.
- [22] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 286–301.
- [23] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng, "A $\hat{2}$ -nets: Double attention networks," *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [24] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.
- [25] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, and S. Yan, "Metaformer is actually what you need for vision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10819–10829.
- [26] S. Maksoud, K. Zhao, C. Peng, and B. C. Lovell, "Scalable Bayesian Deep Learning with Kernel Seed Networks," *arXiv preprint arXiv:2104.09005*, 2021.
- [27] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11534–11542.
- [28] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steinher, D. Keysers, J. Uszkoreit, and others, "MLP-mixer: An all-MLP architecture for vision," *Advances in Neural Information Processing Systems*, vol. 34, pp. 24261–24272, 2021.
- [29] L. Nosrati, M. S. Fazel, and M. Ghavami, "Improving indoor localization using mobile UWB sensor and deep neural networks," *IEEE Access*, vol. 10, pp. 20420–20431, 2022.
- [30] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO Series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.
- [31] H. -S. Choi, J. -H. Kim, J. Huh, A. Kim, J. -W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex u-net," in *Proc. International Conference on Learning Representations*, 2018.
- [32] C. Valentini-Botinhao *et al.*, "Noisy reverberant speech database for training speech enhancement algorithms and TTS models," 2017.
- [33] J. Robert and M. Webbie, "Pydub," 2018.
- [34] ITU-T Recommendation, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P.862*, 2001.
- [35] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, nos. 7-8, pp. 588–601, 2007, Elsevier.
- [36] M. M. Goodwin, "The STFT, sinusoidal models, and speech modification," *Springer Handbook of Speech Processing*, 2008, pp. 229–258.
- [37] Q. Li, J. Wu, Y. Kong, *et al.*, "Speech denoising using only single noisy audio samples," *arXiv e-prints*, 2021, arXiv: 2111.00242.



杨承昊 2022年毕业于厦门大学海洋技术与工程系，获学士学位。目前，他正在厦门大学海洋技术与工程系研究生院攻读硕士学位。他的研究兴趣包括神经网络、音频信号处理和水下声学。



陶毅 1998年毕业于杭州电子工业学院计算机科学与应用专业，获理学学士学位；2008年毕业于厦门大学海洋科学专业，获理学博士学位。现任厦门大学海洋与地球学院海洋物理系助理教授。目前的研究方向是水下声信号处理和水下通信的人工智能方法。



刘敬刘 目前在北京化工大学信息科学与技术学院攻读学士学位。他的研究兴趣包括：无人智能系统、多智能体系统、控制学习以及机器学习。



许肖梅 分别于1982年、1988年和2002年在厦门大学获得海洋物理学学士、硕士和博士学位。她目前是厦门大学应用海洋物理与工程系正教授。她的研究领域包括海洋声学、水下声学遥测和遥控、水下声学通信和信号处理等。

•••