

北京航空航天大学学报
Journal of Beijing University of Aeronautics and Astronautics
ISSN 1001-5965,CN 11-2625/V

《北京航空航天大学学报》网络首发论文

题目：基于多尺度与异构网络的流匹配语音增强
作者：潘嘉杰，贾海蓉，宋郭涛
DOI：10.13700/j.bh.1001-5965.2026.0559
收稿日期：2026-08-07
网络首发日期：2026-11-12
引用格式：潘嘉杰，贾海蓉，宋郭涛. 基于多尺度与异构网络的流匹配语音增强[J/OL]. 北京航空航天大学学报. https://doi.org/10.13700/j.bh.1001-5965.2026.0559



网络首发：在编辑部工作流程中，稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定，且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式（包括网络呈现版式）排版后的稿件，可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定；学术研究成果具有创新性、科学性和先进性，符合编辑部对刊文的录用要求，不存在学术不端行为及其他侵权行为；稿件内容应基本符合国家有关书刊编辑、出版的技术标准，正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性，录用定稿一经发布，不得修改论文题目、作者、机构名称和学术内容，只可基于编辑规范进行少量文字的修改。

出版确认：纸质期刊编辑部通过与《中国学术期刊（光盘版）》电子杂志社有限公司签约，在《中国学术期刊（网络版）》出版传播平台上创办与纸质期刊内容一致的网络版，以单篇或整期出版形式，在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊（网络版）》是国家新闻出版广电总局批准的网络连续型出版物（ISSN 2096-4188，CN 11-6037/Z），所以签约期刊的网络版上网络首发论文视为正式出版。

基于多尺度与异构网络的流匹配语音增强

潘嘉杰，贾海蓉✉，宋郭涛

（太原理工大学 电子信息工程学院，太原 030024）

✉通信作者 贾海蓉 女，博士，教授，硕士生导师。主要研究方向：主要研究方向包括语音信号处理、医疗助听器降噪、微表情识别以及语音处理在人工智能中的应用。E-mail： helenjia722@163.com

摘要 针对扩散方法与语音增强骨干网络的适配性，以及同构双径网络难以有效区分语音时频特性的问题，提出了一个与流匹配高适配的语音增强网络，该网络由多尺度结构与异构双径网络组成。为提取时频多尺度特征，设计了一种条件频率感知时频多尺度融合模块，该模块根据频率特性调整特征权重。为充分挖掘频谱中不同维度特征之间的互补特性，提出了一种异构双径网络，并设计一个频率素核跨组条件特征模块来建模抽象频率特征。此外为了进一步提升流匹配的语音建模能力，提出一种基于条件高斯路径的加权损失函数，将网络输出的向量场转换为对应的估计目标，并引入联合损失进行监督。实验结果表明，提出的方法在参数量、模型性能等方面要高于当前通用的扩散骨干网络，对比基线模型，仅凭借骨干网络便提升了 13.9% 的语音质量感知评估指标，使用联合损失函数后则提到了 27.8%，验证了其有效性。

关键词 语音增强；流匹配；多尺度；双径网络；联合损失函数；扩散模型

中图分类号 TN 912.35 文献标志码 A DOI: 10.13700/j.bh.1001-5965.2026.0559

Flow matching for speech enhancement based on multi-scale and heterogeneous dual-path networks

PAN Jiajie, JIA Hairong✉, SONG Guotao

（College of Electronic Information Engineering, Taiyuan University of Technology, Taiyuan 030024, China）

✉E-mail: helenjia722@163.com

Abstract To address the compatibility issues between diffusion methods and speech enhancement backbone networks, as well as the difficulty of homogeneous dual-path networks in effectively distinguishing speech time-frequency characteristics, a speech enhancement network with high compatibility of flow matching is proposed. This network consists of a multi-scale structure and a dual-path heterogeneous network. To extract multi-scale time-frequency features, a Conditional Frequency-aware Multiscale Fusion module is designed to adjust feature weights based on frequency characteristics. To fully exploit the complementarity of features across different dimensions, the proposed heterogeneous dual-path network incorporates a Prime-kernel Cross-group Conditional Feature Network for modeling abstract frequency features. Moreover, to further improve flow-matching-based speech modeling, a conditional-Gaussian-path weighted loss is proposed, which transforms the predicted vector field into the estimated target and is jointly supervised with an additional loss. Experiments on datasets confirm that the proposed method outperforms diffusion backbones in both parameters and performance. PESQ improves by 13.9% over the baseline with the backbone alone, and by 27.8% when the joint loss is added.

Key words speech enhancement; flow matching; multi-scale; dual-path network; weighted loss; diffusion model

收稿日期：2026-08-07

基金项目：山西省重点研发计划项目（202502010102025）

作者简介：潘嘉杰 男，硕士研究生。主要研究方向：语音增强。

英文基金项目：Shanxi Province Key Research and Development Program (202502010102025)

网络首发时间：2026-11-12 18:18:49 网络首发地址：https://link.cnki.net/urlid/11.2625.V.20261112.1709.001

语音增强作为语音信号处理领域的关键前端技术，旨在从复杂的噪声环境中提取并恢复出干净的语音信号[1]。

相关研究表明，判别式模型通常难以对语音信号在数据空间中的分布进行建模，存在泛化能力不足的问题[2]。上述局限性促使研究者转而采用生成式模型，如生成对抗网络、变分自编码器[3]及扩散去噪模型等。生成式模型的核心优势在于其能够对噪声环境下的干净语音分布进行显式建模，从而获得优于判别式方法的泛化表现。

扩散去噪模型因其优异的生成质量而成为生成式模型中最受关注的范式。Welker 等人[4]将得分匹配应用于语音增强领域，构建了结合均值回归（Ornstein–Uhlenbeck，OU）机制与方差爆炸（variance exploding，VE）噪声调度的随机微分方程（stochastic differential equation，SDE）扩散过程，通过学习条件概率密度的对数梯度，实现生成式语音增强。并在此基础上进一步提出基于得分的生成式语音增强（Score-based Generative Modeling for Speech Enhancement，SGMSE）的改进模型 SGMSE++[5]，该方法的逆扩散过程不再使用纯高斯噪声作为起点，而是以含噪语音与高斯噪声的混合初始状态为起点，实现更高效的生成。Jukić 等人[6]、Zhang 等人[7]与 Nasretdinov 等人[8]相继将薛定谔桥引入生成式语音增强，通过构建连接带噪语音分布与干净语音分布的随机过程，提升了语音增强的性能。Lemercier 等人后续又提出了随机再生模型（stochastic regeneration model，StoRM）[9]，先使用判别式模型得到初步增强结果，随后利用扩散模型进行少步采样，通过随机再生在减少采样步数的同时进一步提升语音生成质量与自然度。Lee 等人提出的 FlowSE[10]，将基于流匹配的扩散模型引入语音增强，该方法采用最优传输（Optimal Transport，OT）条件向量场，学习从噪声分布到干净语音分布的概率路径，并将推理步数减少至 5 步。

相较于噪声条件分数网络（Noise Conditional Score Network，NCSN）[11]逐层下采样处理时频信息的方式，双径结构通过针对不同维度采用不同的特征提取策略，增强模型对长时依赖和局部频谱细节的表达能力。Wang 等人提出的 TF-GridNet[12]交替执行帧内全频带建模与逐频子带时间建模并引入跨帧自注意力机制，从而增强长距离时序依赖建模能力。Saijo 等人针对 DPRNN、DPTNet 等语音处理网络中循环神经网络（recurrent neural network，RNN）并行计算能力受限的问题，提出卷积前馈网络（Convolutional Feed-forward Network，ConvFFN）以代替语音双径 Transformer 中的 RNN 模块[13]。然而语音频谱的时间轴与频率轴各具有不同的物理含义：时间轴主要反映语音信号的动态连续性和时序依赖关系；频率轴主要表征谐波结构、共振峰分布及频带相关特征。因此，根据时域与频域之间的差异设计双径模块具有必要性。Zhuo 等人提出的 MaTSE[14]抛弃传统双径结构，在时间轴与频率轴上分别使用 Mamba 与 Transformer 进行建模，有效避免同构网络造成的特征冗余。

当前大多数扩散模型采用 NCSN++作为骨干网络，并未针对特定扩散模型如流匹配进行设计。因此针对特定扩散过程设计与之适配的骨干网络，是提升模型性能的重要研究方向。

鉴于以上，本文提出一种专用于流匹配方法的语音增强模型，该模型相比传统的 NCSN++拥有更好的泛化性与更低的参数量：1）为解决 NCSN++难以契合频谱时频耦合特性与逐层下采样带来信息损失的问题，提出一个条件频率感知时频多尺度融合模块（Conditional Frequency-aware Multi-scale Fusion Block，CFMF），通过并行部署不同时频感受野的卷积分支，全面捕捉输入特征图的多尺度时频信息，并利用频率感知对其进行加权融合；2）为增强双路径网络对语音时频特征差异的建模能力，本文在瓶颈层采用扩散噪声水平引导的双径异构时频模块（Dual-Path Heterogeneous Time-Frequency Module，DPHTF），并提出了一个与时间 Mamba 异构的频率素核跨组条件特征模块（Prime-kernel Cross-group Conditional Feature Network，PCCFN），用于提取多尺度局部频率模式，通过基于条件信息的跨组动态调制实现组间信息交互，从而提升频率相关性建模能力；3）将扩散模型输出的向量场损失作为主优化目标，在保证流匹配准确性的基础上，引入幅度与相位损失作为辅助约束，以增强频谱结构与相位连续性；进一步结合语音质量感知评估（Perceptual Evaluation of Speech Quality，PESQ）与时域 L1 范数损失作为次级感知约束，对生成语音的主观质量与波形细节进行细化矫正。

1　基于流匹配的语音增强

语音增强中的薛定谔桥通过构建满足端点分布约束的随机过程，实现从带噪语音分布到干净语音的传输。相比之下，流匹配方法则通过学习时间相关向量场，并以其作为监督目标，从而实现数据分布的建模与生成。

基于流匹配（Flow Matching，FM）的生成方法通过学习一个时间连续的向量场 $\mathrm{d}\boldsymbol{x}_t/\mathrm{d}t=\boldsymbol{v}_t(\boldsymbol{x}_t)$ 来构建一个将简单分布逐步变换为目标数据分布的确定性映射过程[10]。

而条件流匹配（Conditional Flow Matching，CFM）利用条件概率路径及其对应的条件向量场构造训练目标，从而学习生成过程所需的时间相关向量场。定义一条连接数据分布与条件先验分布的概率路径 $\boldsymbol{x}_t\sim p_t(\boldsymbol{x})$，并显式构造中间时刻的条件分布 $p_t(\boldsymbol{x}_t\mid\boldsymbol{x}_0,\boldsymbol{y})$。本文参考最优传输条件流匹配（Optimal Transport Conditional Flow Matching，OT-CFM）路径，其具体形式如下所示：

$$\boldsymbol{x}_t=\boldsymbol{\mu}_t(\boldsymbol{x}_0,\boldsymbol{y})+\sigma_{\boldsymbol{x}_t}\boldsymbol{\epsilon},\quad \boldsymbol{\epsilon}\sim\mathcal{N}_{\mathbb{C}}(0,\mathbf{I}) \tag{1}$$

$$\boldsymbol{\mu}_t=(1-t)\boldsymbol{x}_0+t\boldsymbol{y},\quad \sigma_{\boldsymbol{x}_t}=(1-t)\sigma_{\min}+t\sigma_{\max} \tag{2}$$

式中 $\boldsymbol{x}_t$ 为时间 t 的中间状态，$t=T=1$ 为初始时刻，此时 $\boldsymbol{x}_T=\boldsymbol{y}$；$\boldsymbol{\mu}_t$ 与 $\sigma_{\boldsymbol{x}_t}$ 分别为 t 时刻条件概率分布的均值和标准差；$\boldsymbol{\epsilon}$ 为标准复高斯噪声。

CFM 的目标是学习一个时间相关的向量场 $\boldsymbol{v}_\theta(\boldsymbol{x}_t,t,\boldsymbol{y})$ 去拟合目标条件向量场 $\boldsymbol{v}_t(\boldsymbol{x}_t\mid\boldsymbol{x}_0,\boldsymbol{y})$，其中 θ 为神经网络的可训练参数。在给定目标条件概率路径 $p_t(\boldsymbol{x}_t\mid\boldsymbol{x}_0,\boldsymbol{y})=\mathcal{N}\left(\boldsymbol{x}_t;\boldsymbol{\mu}_t\left(\boldsymbol{x}_0,\boldsymbol{y}\right),\sigma_{\boldsymbol{x}_t}^2\mathbf{I}\right)$ 时，可得到其训练目标，如下公式所示：

$$\mathcal{L}_{\mathrm{CFM}}=\mathbb{E}_{t,p_t(\boldsymbol{x}_t\mid\boldsymbol{x}_0,\boldsymbol{y})}[|\boldsymbol{v}_\theta(\boldsymbol{x}_t,t,\boldsymbol{y})-\boldsymbol{v}_t(\boldsymbol{x}_t\mid\boldsymbol{x}_0,\boldsymbol{y})|^2] \tag{3}$$

其中，$\boldsymbol{v}_t(\boldsymbol{x}_t\mid\boldsymbol{x}_0,\boldsymbol{y})$ 可由高斯条件概率路径的均值、标准差及其对时间 t 的导数 $\dot{\boldsymbol{\mu}}_t(\boldsymbol{x}_0,\boldsymbol{y})$ 和 $\dot{\sigma}_{\boldsymbol{x}_t}$ 解析构造，其表达式为：

$$\boldsymbol{v}_t(\boldsymbol{x}_t\mid\boldsymbol{x}_0,\boldsymbol{y})=\frac{\dot{\sigma}_{\boldsymbol{x}_t}}{\sigma_{\boldsymbol{x}_t}}(\boldsymbol{x}_t-\boldsymbol{\mu}_t(\boldsymbol{x}_0,\boldsymbol{y}))+\dot{\boldsymbol{\mu}}_t(\boldsymbol{x}_0,\boldsymbol{y}) \tag{4}$$

采用欧拉离散积分方法对所学习的常微分方程进行数值求解[20]，实现由噪声分布到数据分布的逐步传输，如下公式所示：

$$\boldsymbol{x}_{t_{n-1}}=\boldsymbol{x}_{t_n}-\Delta t_n\cdot\boldsymbol{v}_\theta(\boldsymbol{x}_{t_n},t_n,\boldsymbol{y}),\ \Delta t_n=t_n-t_{n-1}>0 \tag{5}$$

其中，时间节点满足 $t_0<t_1<\cdots<t_{n-1}<t_n<\cdots<t_N$。采样从 $t_N=T=1$ 开始，初始样本 x_T 由先验分布 $\boldsymbol{x}_T\sim\mathcal{N}_{\mathbb{C}}(\boldsymbol{y},\sigma_{\boldsymbol{x}_T}^2\mathbf{I})$ 生成，通过基于所学习的向量场进行离散积分沿时间反向演化更新，该过程对应于确定性的积分形式。

在后续实验中如没有特殊说明，各个参数一律取 $\sigma_{\max}=0.487$，$\sigma_{\min}=0$[10]。

2　本文算法

2.1 模型总体框架

针对上述扩散方法与骨干网络之间的适配性问题，本文提出了一种专用于流匹配方法的语音增强网络，并设计了 CFMF 与 PCCFN 两个模块以提升模型的时频特征建模能力：CFMF 用于缓解无下采样网络中全局上下文感知能力不足的问题；PCCFN 用于适配语音频率轴非平稳、非均匀的分布特性。此外，在训练阶段将网络预测的流匹配向量场转换为中间时刻的目标语音估计，并据此计算辅助损失，以联合优化骨干网络。

其整体网络结构如图 1 所示。其中，图 1(a)中 $\boldsymbol{X}_t$ 为当前时刻的语音信号，$\boldsymbol{Y}$ 为条件信息，在这

里用原始含噪语音表示，output 为模型输出，σ 为噪声水平（等价于时间步 t ）。编码端由残差扩散卷积神经网络（Residual Diffusion Convolutional Neural Network，ResNetDiC）与 CFMF 进行频谱的局部细节优化与层级特征抽象；频谱上下采样仅在频率维度通过卷积实现，以减少时间维度的信息损失。中间层使用时频双径建模的 DPHTF：时间轴前馈网络（Feed-forward Network，FFN）使用 FiLM 改进的双向 Mamba，用于捕捉长距离上下文；频率轴 FFN 采用 PCCFN，用于实现频率维上下文建模与信息交互。解码端采用与编码端对称的 ResNetDiC 与 CFMF，将抽象特征逐步细化；编解码器之间的跳跃连接通过融合对应层级的特征，将条件信息 $\boldsymbol{Y}$ 注入解码器，以引导目标特征的重建。

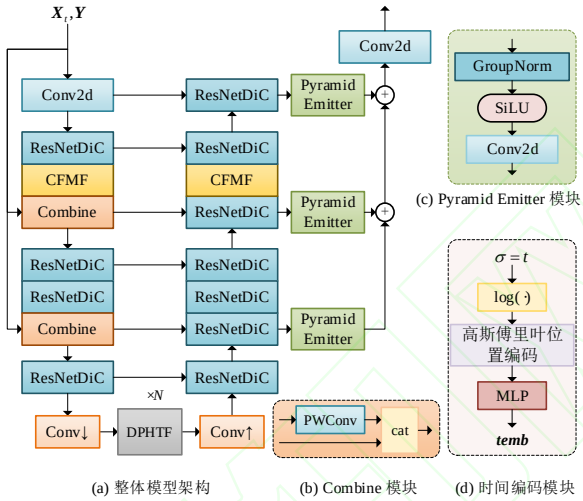


图 1 基于多尺度设计和异构双径网络的语音增强模型

Fig.1 Flow matching for speech enhancement based on multi-scale design and heterogeneous dual-path networks

图 1(b)为 Combine 模块的内部结构，将条件注入分支用点卷积 PWConv 实现通道变换，随后拼接两个输入特征。图 1 (c)为 Pyramid Emitter 模块，由组归一化（Group Normalization，GroupNorm）、SiLU 激活函数与 Conv2d 组成，Pyramid Emitter 作为一个特征处理单元，对当前解码器层的特征进行非线性激活，并通过 Conv2d 将通道维度降维至 4。将重塑后的特征以残差形式累加至金字塔整体输出，在保持时频分辨率不变的同时，补充频谱细节与上下文信息，提升模型输出质量。其中图 1 (d)为模型的噪声水平编码模块：首先对当前时间 t 进行对数变换以实现尺度对齐，随后采用高斯傅里叶位置编码将其映射至高维特征空间，通过多层感知机（Multi-Layer Perceptron，MLP）实现非线性变换和通道映射。

其中网络编解码器参数配置如表 1 所示，分别对应图 1(a)中的各个模块的通道维度变换。

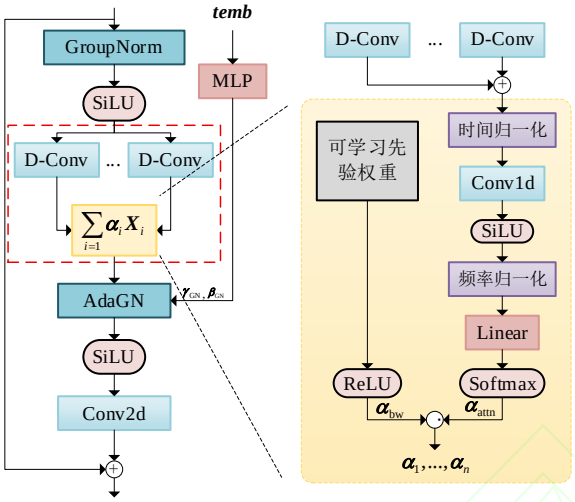
表 1 网络中编码器层与解码器层的参数配置

Table 1 Configuration of channel parameters for the network encoder and decoder layers

模块	通道变换	模块	通道变换
编码器-0: Conv2d	4→64	解码器-0: ResNetDiC	64+64→64
编码器-1: ResNetDiC	64→64	解码器-1: ResNetDiC	64+64→64
编码器-1: CFMF	64→64	解码器-1: CFMF	64→64
编码器-1: Combiner	64+64→64	解码器-1: ResNetDiC	64+64→64
编码器-2: ResNetDiC	64→64	解码器-2: ResNetDiC	64+64→64
编码器-2: ResNetDiC	64→128	解码器-2: ResNetDiC	64→128
编码器-2: Combiner	64+64→64	解码器-2: ResNetDiC	128+128→128
编码器-3: ResNetDiC	128→128	解码器-3: ResNetDiC	128+128→128

2.2 条件频率感知时频多尺度融合模块

为弥补取消逐层下采样后模型在多分辨率特征提取方面的不足，本文采用 CFMF 模块，通过提取适合的时频耦合特征以取代下采用后的多分辨率特征，其详细模型如图 2 所示。



(a) CFMF 模块的详细结构　(b) 频率感知的动静态权重

图 2 CFMF 详细结构
Fig.2 Detailed structure of CFMF

ResNetDiC[15]作为编解码器的基础模块，由两组 GroupNorm、SiLU 激活函数、Conv2d 与残差链接组成，其中第一个 Conv2d 用于通道维度变换，第二个为自适应组归一化（Adaptive Group Normalization，AdaGN），利用时间编码得到的缩放与偏移进行调制，具体形式如下公式所示：

$$\mathrm{AdaGN}(\boldsymbol{X}) = \mathrm{GroupNorm}(\boldsymbol{X}) \odot (\mathbf{1} + \boldsymbol{\gamma}_{\mathrm{GN}}) + \boldsymbol{\beta}_{\mathrm{GN}} \quad (6)$$

式中 $\boldsymbol{\gamma}_{\mathrm{GN}}$ 为缩放参数，$\boldsymbol{\beta}_{\mathrm{GN}}$ 为偏移参数，$\odot$ 为逐元素相乘，$\boldsymbol{X}$ 视作模块函数的输入。

表 2 CFMF模块参数
Tabel 2 CFMF Module Configuration

	空洞卷积 1	空洞卷积 2	空洞卷积 3
卷积核 (F, T)	(3, 3)	(3, 5)	(3, 7)
空洞率 (F, T)	(1, 1)	(2, 4)	(3, 6)

相较于 ResNetDiC，CFMF 模块同样使用两组 GroupNorm、SiLU 激活函数与 Conv2d，但是第一组中的 Conv2d 被替换为多个并行的卷积 D-Conv，每个分支用于捕获不同尺度的时频耦合信息，如图 2 中(a)的红框所示。空洞卷积通过增大采样间隔，在保持参数量和时频分辨率不变的情况下扩大感受野。本文采用不同空洞率，进一步提取多尺度时频上下文信息。其中 CFMF 模块并行空洞卷积参数如表 2 所示。

为了融合上述多分支特征，CFMF 引入频率感知的动静态加权策略，由动态通道注意力机制权重 $\boldsymbol{\alpha}_{\mathrm{attn}}$ 与静态可学习权重 $\boldsymbol{\alpha}_{\mathrm{bw}}$ 构成，其详细结构如图 2 中(b)所示。

通道注意力机制由频率维度的上下文信息来动态生成分支权重。首先，对时频多尺度特征 $\boldsymbol{X}_{\mathrm{C}} = [\boldsymbol{X}_{\mathrm{C}_1}, \boldsymbol{X}_{\mathrm{C}_2}, \boldsymbol{X}_{\mathrm{C}_3}] \in \mathbb{R}^{B\times C\times F\times T}$ 采用时间维度平均池化（$\mathrm{AvgPool}_T$）进行聚合，如下公式所示：

$$\boldsymbol{S} = \mathrm{AvgPool}_T\left(\sum \boldsymbol{X}_{\mathrm{C}_i}\right) \in \mathbb{R}^{B\times C\times F} \quad (7)$$

式中 B、C、F 和 T 分别表示批量大小、通道维度、频率和时间。在不损失频率轴细节的前提下，聚合每个通道沿时间维度的信息，得到表征各通道在不同频点处平均能量分布的特征 $\boldsymbol{S}$。

然后通过一个卷积核为 3 的 Conv1d 捕捉频率局部上下文信息将通道维度降至 $C/16$，并激活生成中间特征 $\boldsymbol{V}$。接着采用频率维度的平均池化（$\mathrm{AvgPool}_F$）对所有频率维度特征进行全局聚合，为每个通道生成紧凑频率特征表示，然后通过线性层（Linear）将通道维度映射至分支数，得到特征 $\boldsymbol{L}$，如下公式所示：

$$\begin{cases} \boldsymbol{V}=\mathrm{SiLU}(\mathrm{Conv1d}(\boldsymbol{S}))\in\mathbb{R}^{B\times(C/16)\times F} \\ \boldsymbol{L}=\mathrm{Linear}(\mathrm{AvgPool}_F(\boldsymbol{V}))\in\mathbb{R}^{B\times C} \end{cases} \tag{8}$$

最后通过 SoftMax 函数将多尺度分支预测的原始分数转换为归一化的注意力权重 $\boldsymbol{\alpha}_{\mathrm{attn}}$。模型能够根据输入特征的动态内容，自适应地调整不同分支的贡献。

静态分支权重 $\boldsymbol{\alpha}_{\mathrm{bw}}$ 由一组可学习参数构成，经 ReLU 激活和归一化后，用于对各多尺度分支进行加权融合。该权重在通道间共享，训练完成后保持固定，以表征各分支的相对重要性。两者相结合得到各个空洞卷积分支的最后权重 $\boldsymbol{\alpha}=\boldsymbol{\alpha}_{\mathrm{attn}}\odot\boldsymbol{\alpha}_{\mathrm{bw}}$，并将该权重与对应的卷积特征融合得到最终的特征输出。

2.3 由 FiLM-BiMamba 与 PCCFN 组成的异构双径网络

DiTs（Diffusion Transformers）[16] 主要应用于图像生成领域，其结构设计对语音时频特征的适配性仍有待提升。为此，本文在 DiTs 的基础上设计了 DPHTF，将 FiLM-Mamba 和 PCCFN 作为时间轴与频率轴的前馈网络，分别用于捕捉时间维的长时依赖与频率维的结构信息，如图 3 所示。

考虑到参数量与泛化性的问题，异构双径网络使用共享多头注意力机制[17]，在图中用 Share 标注。用于噪声水平调制的 MLP-T/F 则不共享参数，目的是学习时域与频域之间的差异，如图 3(b)所示，其中 Shared 表示模块参数共享，如下公式所示：

$$\begin{cases} \mathrm{AdaLN}(\boldsymbol{X})=\mathrm{LayerNorm}(\boldsymbol{X})\odot(\boldsymbol{1}+\boldsymbol{\gamma}_{\mathrm{LN}})+\boldsymbol{\beta}_{\mathrm{LN}} \\ \boldsymbol{X}_{\mathrm{MHSA}}=\boldsymbol{X}_T+\mathrm{MHSA}\big(\mathrm{AdaLN}(\boldsymbol{X}_T)\big)\odot\boldsymbol{gate} \end{cases} \tag{9}$$

其中，$\boldsymbol{X}_T\in\mathbb{R}^{(BF')\times T\times C}$ 为重塑后的时间轴特征，MHSA 为多头自注意力机制，$\boldsymbol{gate}$ 为门控参数。

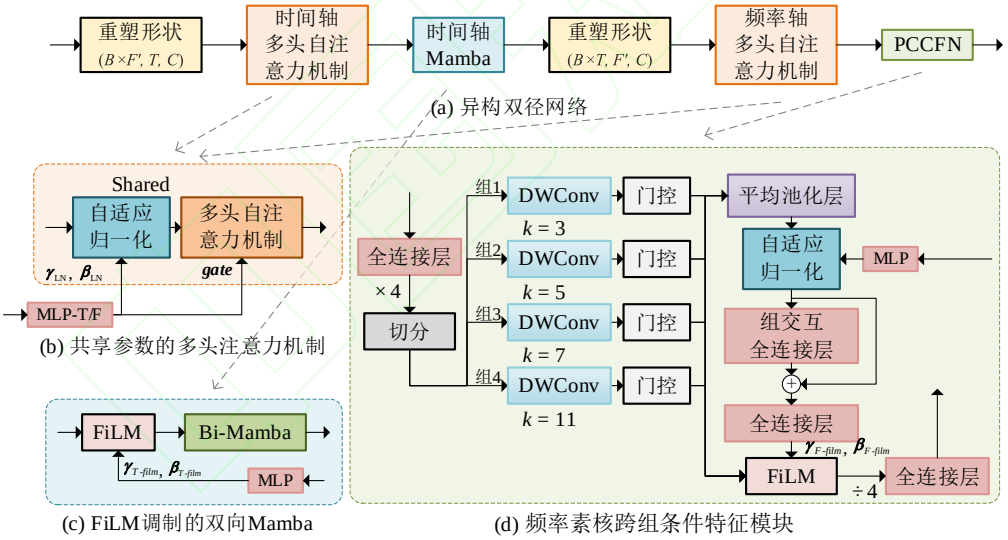


图 3 基于 FiLM-BiMamba 和 PCCFN 的 DPHTF 架构
Fig.3 Architecture of the DPHTF with FiLM-BiMamba and PCCFN.

FFN 在 Transformer 中负责对位置相关的特征做逐点非线性变换，但标准 FFN 不具备输入感知能力，难以动态调整其响应行为。如图 3(c)所示，利用双向 Mamba（Bi-Mamba）动态上下文感知的序列建模能力对时间多头自注意力特征 $\boldsymbol{X}_{\mathrm{MHSA}}$ 进行建模[18]，并采用 FiLM 机制根据噪声水平生成调制参数 $\boldsymbol{\gamma}_{T\text{-}film}$ 和 $\boldsymbol{\beta}_{T\text{-}film}$，得到经时间轴建模后的最终特征 $\hat{\boldsymbol{X}}_{bf',t,c}$，如下公式所示：

$$\hat{\boldsymbol{X}}_{bf',t,c}=\text{Bi-Mamba}(\boldsymbol{X}_{\mathrm{MHSA}}\odot(\boldsymbol{1}+\boldsymbol{\gamma}_{T\text{-}film})+\boldsymbol{\beta}_{T\text{-}film}) \tag{10}$$

考虑到语音频谱沿频率轴的局部相关性与不同频带间的特征差异，本文采用 PCCFN 进行频率轴特征建模。采用多个素数尺寸的卷积核并行处理不同尺度且无遗漏的频率信息，并通过一个上下文感

知的跨组交互机制实现动态融合[19]。

首先将重塑后的频率轴特征 $\boldsymbol{X}_F \in \mathbb{R}^{(BT)\times F'\times C}$ 通过全连接层投影到更高维的空间，随后被划分为 4 个子组 $\boldsymbol{X}_F=[\boldsymbol{X}_{F_1},\boldsymbol{X}_{F_2},\boldsymbol{X}_{F_3},\boldsymbol{X}_{F_4}]$，每个子组 $\boldsymbol{X}_{F_i}$ 对应于一个特定的频率处理路径，其中 F' 为降维后的频率维度。各分支独立地采用结合门控机制的深度可分离卷积进行特征变换，得到特征表示 $\boldsymbol{X}_{bt,f',c}$，其中每个深度分离门控卷积由深度卷积（DWConv）生成特征分支 $\boldsymbol{v}$ 和门控分支 $\boldsymbol{g}$ 与逐点卷积（PWConv）融合通道信息，如下所示：

$$\begin{cases}[\boldsymbol{v},\boldsymbol{g}]=\mathrm{DWConv}(\boldsymbol{X}_{F_i})\\ \boldsymbol{X}_{bt,f',c}=\mathrm{PWConv}(\boldsymbol{v}\odot\mathrm{sigmoid}(\boldsymbol{g}))\end{cases} \tag{11}$$

随后引入跨组交互模块来实现不同频率组之间的信息交换。将所有分组的特征堆叠，然后沿通道维度 C_g 平均池化，如下公式所示：

$$\boldsymbol{r}_{bt,f',g}=\frac{1}{C_g}\sum_{c=1}^{C_g}\boldsymbol{X}_{bt,f',g,c} \tag{12}$$

通过特征聚合，得到各组的紧凑特征表示 $\boldsymbol{R}=[\boldsymbol{r}_{bt,f',g}]\in\mathbb{R}^{(BT)\times F'\times 4}$。

利用时间编码对聚合后的组特征 $\boldsymbol{R}$ 施加条件归一化，使模型能根据噪声水平动态调整组特征的统计特性，增强其对噪声水平的感知能力。为增强通道间的信息交互，进一步通过全连接层对调制后的组特征进行变换得到 $\tilde{\boldsymbol{R}}$，如下所示：

$$\tilde{\boldsymbol{R}}=(\boldsymbol{1}+\boldsymbol{W}_{cg})\odot\mathrm{AdaLN}(\boldsymbol{R}) \tag{13}$$

式中 $\boldsymbol{W}_{cg}\in\mathbb{R}^{4\times4}$ 表示可学习的跨组交互矩阵。

组特征 $\tilde{\boldsymbol{R}}$ 通过全连接层得到组间调制参数 $\boldsymbol{\gamma}_{F\text{-}film}$ 和 $\boldsymbol{\beta}_{F\text{-}film}$，随后被用于对初始分组卷积输出 $\boldsymbol{X}_{bt,f',c}$ 做 FiLM 调制，进一步促进不同频率组之间的信息交互，使各组特征根据其他组的聚合信息及噪声水平进行自适应调整，其计算过程为：

$$\begin{cases}[\boldsymbol{\gamma}_{F\text{-}film},\boldsymbol{\beta}_{F\text{-}film}]=\boldsymbol{W}_{film}\hat{\boldsymbol{R}}+\boldsymbol{b}_{film}\\ \hat{\boldsymbol{X}}_{bt,f',c}=\boldsymbol{X}_{bt,f',c}\odot(\boldsymbol{1}+\boldsymbol{\gamma}_{F\text{-}film})+\boldsymbol{\beta}_{F\text{-}film}\end{cases} \tag{14}$$

式中 $\boldsymbol{W}_{film}$ 表示可学习的全连接层，$\boldsymbol{b}_{film}$ 为初始化为 0 的偏移矩阵。

最后通过全连接层将特征投影回原始维度空间。设置双径异构时频模块数量为 4。

2.4 条件流匹配联合损失

在流匹配模型中，骨干网络输出的是向量场 $\boldsymbol{v}_\theta$，而语音增强任务中的联合损失函数所需的是增强后的语音信号 $\boldsymbol{x}_\theta$，如下公式所示：

$$\begin{cases}\boldsymbol{v}_\theta=F_\theta(\boldsymbol{x}_t,t,\boldsymbol{y})\\ \boldsymbol{x}_t=(1-t)\boldsymbol{x}_\theta+t\boldsymbol{y}+((1-t)\sigma_{\min}+t\sigma_{\max})\boldsymbol{\epsilon}\\ \boldsymbol{v}_\theta=-\boldsymbol{x}_\theta+\boldsymbol{y}+(\sigma_{\max}-\sigma_{\min})\boldsymbol{\epsilon}\end{cases} \tag{15}$$

式中，$\boldsymbol{y}$ 为条件信息，也就是初始含噪语音；F_θ 为骨干网络，其输出为条件向量场 $\boldsymbol{v}_\theta$。

为了解决难以将辅助损失函数加入流匹配语音增强训练的问题，将骨干网络输出当前噪声条件下的向量场 $\boldsymbol{v}_\theta$ 与对真实样本的最优估计 $\boldsymbol{x}_\theta$ 进行转换，可以推导出以下公式：

$$\boldsymbol{x}_\theta=\frac{\sigma_{\max}-\sigma_{\min}}{\sigma_{\max}}\boldsymbol{x}_t+\frac{\sigma_{\min}}{\sigma_{\max}}\boldsymbol{y}-\frac{(1-t)\sigma_{\min}+t\sigma_{\max}}{\sigma_{\max}}\boldsymbol{v}_\theta \tag{16}$$

为准确重建目标语音的频谱幅度，本文引入幅度谱损失作为优化目标，其计算公式如下：

$$\mathcal{L}_{\text{Mag.}} = \mathbb{E}_{\boldsymbol{X}_\theta, \boldsymbol{X}_0}[\| |\boldsymbol{X}_\theta| - |\boldsymbol{X}_0| \|_2^2] \tag{17}$$

式中$|\boldsymbol{X}_\theta|$为增强语音频谱幅度，$|\boldsymbol{X}_0|$为干净语音频谱幅度。

将复频谱相位定义为$\phi(\boldsymbol{X}) = \text{atan2}(\boldsymbol{X}_i, \boldsymbol{X}_r)$，其中$\boldsymbol{X}_i$是语音复频谱虚部，$\boldsymbol{X}_r$是复频谱实部。同时为了解决相位差之间的周期问题，引入相位包裹函数$\mathcal{W}(\phi(\boldsymbol{X})) = \text{atan2}(\sin\phi(\boldsymbol{X}), \cos\phi(\boldsymbol{X}))$，将相位差限制在$(-\pi, \pi]$。

定义瞬时相位差：$\Delta\phi = \mathcal{W}(\phi(\boldsymbol{X}_\theta) - \phi(\boldsymbol{X}_0))$，此时可得到瞬时相位（instantaneous phase，IP）损失函数，公式如下所示：

$$\mathcal{L}_{\text{IP}} = \mathbb{E}_{\boldsymbol{X}_\theta, \boldsymbol{X}_0}[1 - \cos(\Delta\phi)] \tag{18}$$

进一步地，将时间方向上的相位差定义为：$\Delta_t\phi(\boldsymbol{X}) = \mathcal{W}(\phi(\boldsymbol{X})_{f,t+1} - \phi(\boldsymbol{X})_{f,t})$，表示为沿时间轴的微分算子，瞬时角频率（instantaneous angular frequency，IAF）损失函数如下所示：

$$\mathcal{L}_{\text{IAF}} = \mathbb{E}_{\Delta_t(\boldsymbol{X}_\theta, \boldsymbol{X}_0)}[1 - \cos(\Delta_t\phi(\boldsymbol{X}_\theta) - \Delta_t\phi(\boldsymbol{X}_0))] \tag{19}$$

同时，将频率方向上的相位差定义为：$\Delta_f\phi(\boldsymbol{X}) = \mathcal{W}(\phi(\boldsymbol{X})_{f+1,t} - \phi(\boldsymbol{X})_{f,t})$，表示为沿频率轴的微分算子，群延迟（group delay，GD）损失函数如下所示：

$$\mathcal{L}_{\text{GD}} = \mathbb{E}_{\Delta_f(\boldsymbol{X}_\theta, \boldsymbol{X}_0)}[1 - \cos(\Delta_f\phi(\boldsymbol{X}_\theta) - \Delta_f\phi(\boldsymbol{X}_0))] \tag{20}$$

总相位损失定义为瞬时相位损失、瞬时角频率损失与群延迟损失之和。

$$\mathcal{L}_{\text{Pha.}} = \mathcal{L}_{\text{IP}} + \mathcal{L}_{\text{IAF}} + \mathcal{L}_{\text{GD}} \tag{21}$$

$\mathcal{L}_{\text{Pesq}}$损失函数基于 PESQ 评分进行测量，采用 torch_pesq 库中的 PesqLoss 函数进行实现[21]。

为提升波形重建能力，本文引入时域损失函数，在预测频谱图重建的波形与原始干净波形之间计算 L1 距离，其定义如下：

$$\mathcal{L}_{\text{Dom.}} = \mathbb{E}_{\boldsymbol{X}_\theta^{\text{td}}, \boldsymbol{X}_0^{\text{td}}}[\| \boldsymbol{X}_\theta^{\text{td}} - \boldsymbol{X}_0^{\text{td}} \|_1] \tag{22}$$

式中$\boldsymbol{X}_\theta^{\text{td}}$为时域增强语音信号，$\boldsymbol{X}_0^{\text{td}}$为时域干净语音信号。

基于以上损失描述，构建了一种条件流匹配联合损失，其整体表达式如下所示：

$$\begin{aligned}\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CFM}} + \lambda_{\text{Mag}}\mathcal{L}_{\text{Mag.}} + \lambda_{\text{Pha}}\mathcal{L}_{\text{Pha.}} \\ + \lambda_{\text{Pesq}}\mathcal{L}_{\text{Pesq}} + \lambda_{\text{Dom}}\mathcal{L}_{\text{Dom.}}\end{aligned} \tag{23}$$

式中各损失函数权重为：$\lambda_{\text{Mag}} = 0.1$，$\lambda_{\text{Pha}} = 0.01$，$\lambda_{\text{Pesq}} = 0.001$，$\lambda_{\text{Dom}} = 0.01$。

该方法弥合了流匹配目标与语音增强目标之间的优化差距，为语音域先验知识和感知约束加入流匹配框架提供了一种可行的方案。

3 实验与结果分析

3.1 实验设置

为了充分评估本文提出方法的性能，本章选取两个数据集，分别是 VoiceBank+Demand（VB-DMD）[22]与 LibriSpeech+NoiseX92。其中 VB-DMD 数据集是一个公共数据集，由 30 名说话人（15 男 15 女）提供的 12396 条语音构成，训练集与验证集中的含噪语音的信噪比设置为 0、5、10、15 dB，测试数据设置信噪比为 2.5、7.5、12.5、17.5 dB。在训练数据中将说话人 p226 和 p287 用作验证集（820 条），p232 和 p257 用作测试集（824 条），其余说话人用于训练集（10752 条）。LibriSpeech+NoiseX92 是一个自建的数据集，用于测试模型方法在低信噪比下的性能与泛化性。训练集的干净语音取自 LibriSpeech[23]中的一个子集 train-clean-100，由 251 名说话人构成，验证集选用 dev-clean，测试集选用 test-clean。数据集中的噪声取自 NoiseX92[24]中的 15 种噪声，将 11 种噪声随机应用于训练集与验证集，信噪比随机设置为[-10 dB, 10 dB]之间的 21 个整数，然后用剩下的 4 种噪

声创建测试集，信噪比设置为[-5 dB, -2.5 dB, 0 dB, 2.5 dB, 5 dB]。

在训练阶段，所有音频均被切分为 2 秒的片段，并将采样率统一为 16 kHz、位数为 16 bit。本实验采用短时傅里叶变换（Short-time Fourier Transform，STFT）与汉明窗，快速傅里叶变换点数和窗长均为 510，帧移为 128 个采样点，每个样本包含 256 帧。据此提取的语音频谱图具有 $F=256$ 的频率分量和 $T=256$ 的时间帧。在优化器方面使用了 Adam，初始学习率设置为 1e-4，其他超参数均采用 PyTorch 默认值（$\beta_1=0.9$，$\beta_2=0.999$）。为了提高模型的泛化性能并获得更稳定的最终模型权重，在训练过程中采用了指数移动平均策略来更新模型参数，设置衰减因子 α=0.999。所有模型训练 400k 步。

3.2 性能对比

在 VB-DMD 数据集上与多个判别式与生成式模型进行对比。采用 PESQ[25]、尺度不变信号失真比（Scale-invariant Signal-to-distortion Ratio，SI-SDR）[26]、扩展短时客观可懂度（Extended Short-time Objective Intelligibility，ESTOI）[27]等作为评估指标。各个对比模型的数据从现公开的论文中获取，结果如表 3 所示。其中，D 代表判别式模型，G 代表生成式模型，Ours(w/o Loss) 表示仅采用流匹配损失训练的模型，而 Ours(w Loss)表示采用所提出加权联合损失训练的完整模型。

表 3 提出模型在VB-DMD数据集上的对比实验

Table 3 Comparative experiments of the proposed model on the VB-DMD dataset

模型	Type	PESQ	SI-SDR	ESTOI	参数量/M
Noisy	—	1.97	8.4	0.79	—
MetricGAN+[28]	D	3.13	8.5	0.83	—
CleanMel[29]	D	2.89	—	—	7.2
CDiffuSE[30]	G	2.58	12.4	0.79	—
SGMSE++	G	2.93	17.3	0.87	65.6
StoRM[9]	D+G	2.93	18.8	0.88	27.8+27.5
SBVE[31]	G	2.91	19.4	0.88	65.6
FlowSE[10]	G	3.12	18.9	0.88	65.6
TM-DBA-M[32]	G	3.17	19.4	0.89	10.44
Ours(w/o Loss)	G	3.28	19.4	0.89	7.7
Ours(w Loss)	G	3.44	19.3	0.89	7.7

实验结果表明，未引入加权联合损失时，本文模型的 PESQ 较基线 FlowSE 提高 13.9%；引入该损失后，相对基线的提升幅度达到 27.8%，PESQ 到达了 3.44。上述结果说明，骨干网络的改进已有效提升语音增强性能，而加权联合损失进一步改善了增强语音的感知质量。

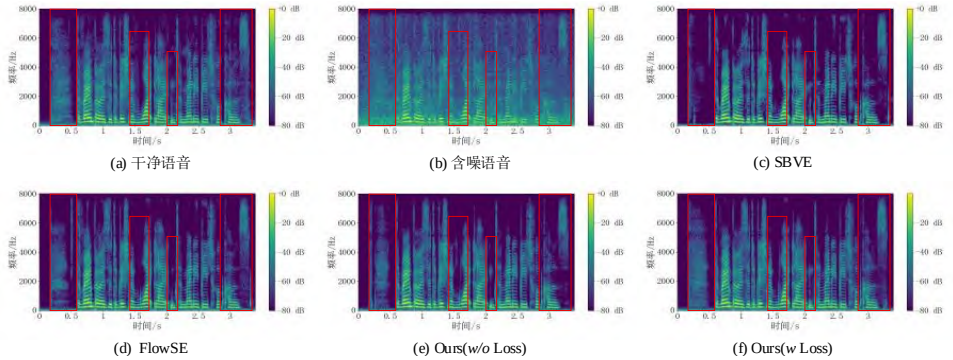
表 4 骨干网络参数量与计算量对比

Table 4 Comparison of parameter and computational complexity in backbone networks

骨干网络	参数量/M	计算量（FLOPs/G）
NCSN++	65.6	133
Ours	7.7	146

表 4 对比了各骨干网络的参数量与计算量。从中观察到本文所提出骨干网络的参数量相较于 NCSN++大幅降低。由于该方法仅在频率维下采样，在时间维保持全分辨率以保留时序细节，其单次前向计算量相较 NCSN++轻微增加。综合而言，本文提出的骨干网络以计算量略微提升为代价，换取了参数量的显著压缩与生成质量的提升。

此外绘制测试集中的样本 p257_077.wav 的 STFT 可视化如图 5 所示，通过与各模型进行比较，从红框位置可见提出模型的增强语音频谱与干净语音频谱具有更高的一致性，能够更有效地保留语音细节并抑制噪声干扰。



(a) 干净语音 (b) 含噪语音 (c) SBVE

(d) FlowSE (e) Ours(*w/o* Loss) (f) Ours(*w* Loss)

图 4 各模型在 p257_007.wav 上的 STFT 可视化图

Fig.4 The STFT visualizations of each model on p257_007.wav

为了验证本文提出模型在低信噪比场景下的泛化能力，选择 LibriSpeech+NoiseX92 数据集分别在采用方差爆炸噪声调度的薛定谔桥（Schrödinger bridge with variance exploding，SBVE）和 OT-CFM 框架下，对比采用 NCSN++与所提骨干网络的模型，并与经典扩散语音增强模型进行比较。实验结果如表 5 所示。其中：实验 1：将 NCSN++作为骨干网络在 OT-CFM 模型下运行；实验 2：将 NCSN++作为骨干网络在 SBVE 模型下运行；实验 3：将提出模型作为骨干网络在 OT-CFM 模型下运行；实验 4：将提出模型作为骨干网络在 SBVE 模型下运行。

表 5 在LibriSpeech+NoiseX92数据集上的对比实验

Table 3 Comparative experiments of the proposed model on the LibriSpeech+NoiseX92

	PESQ					ESTOI				
SNR/dB	-5	-2.5	0	2.5	5	-5	-2.5	0	2.5	5
Noisy	1.09	1.13	1.19	1.26	1.35	0.48	0.54	0.60	0.66	0.71
实验 1	1.39	1.51	1.65	1.81	1.99	0.68	0.73	0.77	0.80	0.84
实验 2	1.36	1.46	1.57	1.71	1.87	0.71	0.75	0.79	0.82	0.85
实验 3	1.60	1.72	1.85	2.01	2.17	0.70	0.75	0.78	0.81	0.84
实验 4	1.30	1.39	1.43	1.50	1.57	0.71	0.72	0.74	0.76	0.78
SGMSE++	1.37	1.47	1.58	1.74	1.89	0.71	0.74	0.78	0.81	0.84
StoRM	1.39	1.45	1.61	1.76	1.91	0.71	0.74	0.79	0.81	0.84
Ours(*w* Loss)	1.77	1.88	2.01	2.16	2.35	0.74	0.78	0.80	0.83	0.86

通过对比提出模型与 NCSN++在 SBVE 与 OT-CFM 扩散方法上的实验结果，可见提出模型在较低信噪比条件下具有优于桥方法的泛化能力，其在单边约束扩散（OT-CFM）下性能表现优异，而在双端约束扩散（SBVE）下，提出模型与 NCSN++相比，其泛化性更差，难以释放全部性能。该表现源于所提骨干网络设计专为单边 OT-CFM 路径优化：其多尺度、双径结构与单端约束下的监督信息高度匹配，因而能高效逼近理论轨迹；而在 SBVE 双端耦合约束下，其监督信息利用效率下降，导致学习到的概率流偏离理论最优传输路径，PESQ 随之下降。说明提出模型在 OT-CFM 方法下的适配性更好，具有更强的语音生成建模能力。

通过对比表 5 中 SGMSE++、StoRM 等扩散经典模型与提出方法的性能对比，观察到本文模型在 ESTOI 指标上的优势相对有限，但在 PESQ 指标上取得了显著领先的性能，这说明所提方法能够更有效地改善语音质量和听觉感知效果。

3.3 消融实验

为了评估本文所提出网络模型在多种扩散方法下的性能，本节使用 VB-DMD 数据集来作为消融实验的基线数据集，采用 OUVE、SBVE、OT-CFM 等扩散方法，其中 OUVE 的参数设置与论文[28]一致。

表 6 不同扩散方法的对比实验

Table 6 Comparative experiments of the proposed model using different diffusion models

扩散模型	采样步数	PESQ	SI-SDR	ESTOI
Noisy	—	1.97 ±0.75	8.4 ±5.6	0.79 ±0.15
OUVE[28]	30	3.05 ±0.78	18.3 ±4.6	0.87 ±0.09
SBVE[28]	50	3.09 ±0.79	17.2 ±5.2	0.89 ±0.09
OT-CFM	5	3.28 ±0.70	19.4 ±3.3	0.89 ±0.09

表 6 对比骨干网络在三种扩散方法下的性能指标，结果表明提出的骨干网络在 OT-CFM 下表现最佳。该方法在仅需 5 步采样的情况下，取得 3.28 的 PESQ 以及 19.4 dB 的 SI-SDR，兼顾了性能与效率。

为了验证模型中各个模块的有效性，本节通过与使用较多的代表性模块 ResNetDiC、GRU-FFN[33]、Bi-Mamba 进行消融实验，来量化评估本文方法在语音增强任务中的有效性与鲁棒性，其中“-”代表除去该模块，“+”表示添加该模块，结果如表 7 所示。

表 7 基于OT-CFM模型的CFMF和PCCFN模块消融实验

Table 7 Ablation experiments on the CFMF and PCCFN modules within the OT-CFM-based model

模型	PESQ	SI-SDR	ESTOI	参数量/M
Noisy	1.97 ±0.75	8.4 ±5.6	0.79 ±0.15	——
NCSN++(OT-CFM)	3.12 ±0.70	18.9 ±3.3	0.88 ±0.09	65.6
Ours(w/o Loss)	3.28 ±0.70	19.4 ±3.3	0.89 ±0.09	7.7
- CFMF + ResNetDiC	3.21 ±0.70	19.7 ±3.2	0.88 ±0.09	6.8
- PCCFN + GRU-FFN	3.16 ±0.68	19.1 ±3.5	0.88 ±0.09	9.8
- PCCFN + Bi-Mamba	2.97 ±0.76	18.2 ±3.3	0.87 ±0.10	8.8

观察到引入 CFMF 虽然增加 0.9M 的参数量，但显著提升了 PESQ 与 ESTOI 指标。与 GRU-FFN 相比，PCCFN 在减少参数量的同时取得了更好的语音增强性能；与 Bi-Mamba 相比，PCCFN 能够更好地重建频谱细节，表明其在本文双路径结构中更适合频率轴特征建模。上述结果从性能对比的角度验证了 CFMF 与 PCCFN 对整体模型性能的提升作用。

为验证 CFMF 中卷积核尺寸配置的有效性，进一步设置不同卷积核尺寸组合的对比实验，结果如表 8 所示。

表 8 各个卷积组合性能验证

Table 8 Performance Validation of Various Convolution Combinations

	频率多尺度			时频多尺度			时间多尺度		
	D1	D2	D3	D1	D2	D3	D1	D2	D3
卷积核 (F, T)	(3, 3)	(5, 3)	(7, 3)	(3, 3)	(3, 5)	(3, 7)	(3, 3)	(3, 5)	(3, 7)
空洞率 (F, T)	(1, 1)	(2, 1)	(3, 1)	(1, 1)	(2, 4)	(3, 6)	(1, 1)	(1, 4)	(1, 6)
PESQ	3.16 ±0.70			3.28 ±0.70			3.23 ±0.70		
SI-SDR	19.5 ±3.3			19.4 ±3.3			19.7 ±3.2		
ESTOI	0.88 ±0.09			0.89 ±0.09			0.88 ±0.09		

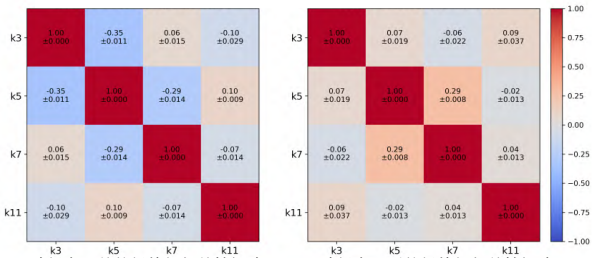
从表中可以观察到，当时间维采用尺寸为 3 的卷积核、频率维采用多尺度卷积核时，模型性能最差；仅在时间维采用多尺度卷积核时，性能有所改善；在时间维与频率维均采用多尺度卷积核时，PESQ 和 ESTOI 取得最优结果。进一步证明本文提出的频率感知与时频多尺度方法有效。

为验证所提出的跨组条件调制对不同感受野组间信息交互的作用，本节对跨组交互模块前后的组间相关性矩阵进行可视化分析。分别保存交互前的分组特征 $\boldsymbol{X}_{bt,f',c}$ 与经过跨组交互后的输出特征 $\hat{\boldsymbol{X}}_{bt,f',c}$，并将每个张量展平后计算不同组之间的皮尔逊相关系数。其中，不同组分别对应不同的素核卷积尺寸，例如 k3 表示卷积核尺寸为 3 的一维空洞卷积，矩阵元素为测试样本上的均值，并标注±95%的置信区间。

如图 5(a)所示，在跨组交互之前，各分组特征虽已表现出一定的正交性，如 k3 与 k7 为 0.06，但存在明显的负相关：k3 与 k5 为-0.35；k5 与 k7 为-0.29，表明各分组虽能提取不同的频率模式信息，但组间仍存在一定冗余与负相关冲突。

如图 5(b)所示，经过跨组模块处理后，上述负相关冲突得到有效抑制。k3 与 k5 由-0.35 变为 0.07，k7 与 k11 为 0.04，k5 与 k11 为-0.02；个别组对如 k5 与 k7 由-0.29 转为 0.29。对角线外平均绝对相关系数由交互前的 0.165 降至交互后的 0.111。上述结果表明，经过跨组模块调制后各组别之间

存在的负相关冲突得到有效抑制，模型展现出更为理想的解耦特性。



(a) 跨组交互前的组特征相关性矩阵 (b) 跨组交互后的组特征相关性矩阵

图 5 PCCFN 跨组动态交互模块的性能验证

Fig.5 Performance validation of the PCCFN cross-group dynamic interaction module

此外，通过移除 PCCFN 中的分组调制模块（仅保留 FiLM 噪声水平调制，由于输入 FiLM 的通道维度不同，导致参数量的增加）进行模块内消融实验，结果如表 9 所示。结合图 5，观察到本文所提出的模块并非简单地独立处理不同的组，而是能够在不同感受野分支之间建立有效的信息交互，从而增强特征的互补性与表达能力，与表中客观语音增强指标提升的结果保持一致。

表 9 PCCFN跨组动态交互性能验证

Table 9 Performance Validation of the PCCFN Cross-Group Dynamic Interaction Module

	PESQ	SI-SDR	ESTOI	参数量
组交互	3.28 ±0.70	19.4 ±3.3	0.89 ±0.09	7.7
无组交互	3.20 ±0.68	19.4 ±3.4	0.88 ±0.09	8.1

为了验证联合损失函数的有效性，本文开展了损失函数消融实验，结果如表 10 所示。其中 −Mag、−Pha、−Pesq、−Dom 分别表示将幅度损失、相位损失、PESQ 损失与时域损失项置零。

表 10 损失函数消融验证

Table 10 Loss function ablation validation

	PESQ	SI-SDR	ESTOI
Ours(w/o loss)	3.28 ±0.70	19.4 ±3.3	0.89 ±0.09
Ours(w loss)	3.44 ±0.66	19.3 ±3.3	0.89 ±0.09
-Mag	3.41 ±0.65	19.1 ±3.2	0.89 ±0.09
-Pha	3.40 ±0.64	19.1 ±3.3	0.89 ±0.09
-Pesq	3.32 ±0.69	19.3 ±3.1	0.89 ±0.09
-Dom	3.40 ±0.65	19.1 ±3.2	0.89 ±0.09

从表 10 观察到，PESQ 损失的加入使该指标提升最为明显，但 PESQ 指标并不能完全反映语音质量的真实改善，并会略微牺牲 SI-SDR；而幅度、相位与时域损失则更有利于提升增强语音的整体质量。

4 结 论

为了解决扩散模型中骨干网络与扩散架构的匹配不足，神经网络难以通过时频解耦特征恢复干净语音的问题，本文提出一种基于多尺度设计和异构双径网络的流匹配语音增强算法。

1）通过 CFMF 模块提取不同时频感受野的卷积特征，利用频率感知机制对这些多尺度特征进行加权融合，有效解决感受野不足、难以契合频谱的时频耦合特性的问题，同时也避免了时间下采样带来的信息损失。

2）提出结合 FiLM-BiMamba 与 PCCFN 的异构双径网络，并在 PCCFN 中通过基于条件信息的跨组动态调制实现组间信息交互，进一步提升了模型的表达能力包括复杂频率相关性的建模能力。

3）采用条件流匹配向量场损失作为主优化目标以保障精度，引入幅度与相位损失，并辅以 PESQ 损失与时域 L1 损失作为约束，有效增强了频谱结构和相位连续性，并对生成语音的波形细节进行精细校正。

4）为验证所提方法的有效性，在 VB-DMD 与 LibriSpeech+NoiseX92 数据集上进行测试，结果表明该方法优于现有的扩散语音增强，在性能指标上表现优异。

尽管本研究提出了一个适用于流匹配的骨干网络，但仍存在一定的不足之处。当前模型在参数效率与计算效率方面仍有不足，这为其在资源受限的移动设备上的部署带来了制约，有待进一步压缩与加速。在实时性方面，该模型为非因果模型，难以满足实时推理的需要，距离流式部署尚有距离。未来将引入多模态视听数据以进一步提升模型性能，并在此基础上探究模型的轻量化和实时性优化方案，推动其在高保真、高效率语音重建任务中的应用。

参考文献

[1] JENIFA G,SAI LIKHITA R M,VINEELA M,et al.Integrating advanced convolutional and recurrent architectures for speech enhancement[C]//2025 6th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks(ICICV).Tirunelveli,India:IEEE,2025,920–928.

[2] XIE Y,TAN Z H.A survey of deep learning for complex speech spectrograms[J].Speech Communication,2025,175:103319.

[3] BANDO Y,MIMURA M,ITOYAMA K,et al.Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization[C]//IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP).Calgary,Canada:IEEE,2018,716–720.

[4] WELKER S,RICHTER J,GERKMANN T.Speech enhancement with score-based generative models in the complex STFT domain[C]//Annual Conference of the International Speech Communication Association(Interspeech).Incheon,Korea:ISCA,2022,2928–2932.

[5] RICHTER J,WELKER S,LEMERCIER J M,et al.Speech enhancement and dereverberation with diffusion-based generative models[J].IEEE/ACM Transactions on Audio Speech and Language Processing,2023,31:2351–2364.

[6] JUKIĆ A,KOROSTIK R,BALAM J,et al.Schrödinger bridge for generative speech enhancement[C]//Annual Conference of the International Speech Communication Association(Interspeech).Kos,Greece:ISCA,2024,1175–1179.

[7] ZHANG H,LI G,WU P,et al.SB-SENet:diffusion model based on Schrödinger bridge for speech enhancement[J].Applied Acoustics,2025,236:110742.

[8] NASRETDINOV R,KOROSTIK R,JUKIĆ A.Robust speech recognition with Schrödinger bridge-based speech enhancement[C]//IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP).Hyderabad,India:IEEE,2025,1–5.

[9] LEMERCIER J M,RICHTER J,WELKER S,et al.StoRM:a diffusion-based stochastic regeneration model for speech enhancement and dereverberation[J].IEEE/ACM Transactions on Audio Speech and Language Processing,2023,31:2724–2737.

[10] LEE S,CHEONG S,HAN S,et al.FlowSE:flow matching-based speech enhancement[C]//IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP).Hyderabad,India:IEEE,2025,1–5.

[11] SONG Y,ERMON S.Generative modeling by estimating gradients of the data distribution[C]//Advances in Neural Information Processing Systems.Vancouver,BC,Canada:Neural Information Processing Systems Foundation,2019,32.

[12] WANG Z Q,CORNELL S,CHOI S,et al.TF-GridNet:integrating full-and sub-band modeling for speech separation[J].IEEE/ACM Transactions on Audio Speech and Language Processing,2023,31:3221–3236.

[13] LE X,CHEN H,CHEN K,et al.DPCRN:dual-path convolution recurrent network for single channel speech enhancement[C]//Annual Conference of the International Speech Communication Association(Interspeech).Brno,Czechia:ISCA,2021,821–825.

[14] ZHOU H,ZHOU Y,ZHAO Y,et al.MaTSE:a hybrid Mamba-Transformer model for monaural speech enhancement[J].Speech Communication,2026,180:103395.

[15] TIAN Y,HAN J,WANG C,et al.DiC:rethinking Conv3x3 designs in diffusion models[C]//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.Nashville,TN,USA:IEEE,2025,2469–2478.

[16] PEEBLES W,XIE S.Scalable diffusion models with transformers[C]//Proceedings of the IEEE International Conference on Computer Vision.Paris,France:IEEE,2023,4172–4182.

[17] KÜHNE N L,JENSEN J,ØSTERGAARD J,et al.MambAttention:Mamba with multi-head attention for generalizable single-channel speech enhancement[J].IEEE/ACM Transactions on Audio Speech and Language Processing,2026,34:820–833.

[18] CHAO R,CHENG W H,QUATRA M L,et al.An investigation of incorporating Mamba for speech enhancement[C]//Proceedings of 2024 IEEE Spoken Language Technology Workshop,SLT 2024.Macao,China:IEEE,2024,302–308.

[19] LIN Z,WANG J,LI R,et al.PrimeK-Net:multi-scale spectral learning via group prime-kernel convolutional neural networks for single channel speech enhancement[C]//IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP).Hyderabad,India:IEEE,2025,1–5.

[20] GONZALEZ P,TAN Z H,ØSTERGAARD J,et al.Diffusion-based speech enhancement in matched and mismatched conditions using a Heun-based sampler[C]//IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP).Seoul,Korea:IEEE,2024,10431–10435.

[21] MARTIN-DONAS J M,GOMEZ A M,GONZALEZ J A,et al.A deep learning loss function based on the perceptual evaluation of the speech quality[J].IEEE Signal Processing Letters,2018,25(11):1680–1684.

[22] VALENTINI-BOTINHAO C,WANG X,TAKAKI S,et al.Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech[C]//9th ISCA Speech Synthesis Workshop,SSW 2016.Sunnyvale,CA,USA:ISCA,2016,146–152.

[23] PANAYOTOV V,CHEN G,POVEY D,et al.Librispeech:an ASR corpus based on public domain audio books[C]//IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP).Brisbane,Australia:IEEE,2015,5206–5210.

[24] VARGA A,STEENEKEN H J M.Assessment for automatic speech recognition:II.NOISEX-92:a database and an experiment to study the effect of additive noise on speech recognition systems[J].Speech Communication,1993,12(3):247–251.

[25] RIX A W,BEERENDS J G,HOLLIER M P,et al.Perceptual evaluation of speech quality(PESQ):a new method for speech quality assessment of telephone networks and codecs[C]//IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP).Salt Lake City,USA:IEEE,2001,749–752.

[26] LE ROUX J,WISDOM S,ERDOGAN H,et al.SDR:half-baked or well done?[C]//IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP).Brighton,UK:IEEE,2019,626–630.

[27] JENSEN J,TAAL C H.An algorithm for predicting the intelligibility of speech masked by modulated noise maskers[J].IEEE/ACM Transactions on Audio Speech and Language Processing,2016,24(11):2009–2022.

[28] FU S W,YU C,HSIEH T,et al.MetricGAN+:an improved version of MetricGAN for speech enhancement[C]//Annual Conference of the International Speech Communication Association(Interspeech).Brno,Czech Republic:ISCA,2021,201–205.

[29] SHAO N,ZHOU R,WANG P,et al.CleanMel:mel-spectrogram enhancement for improving both speech quality and ASR[J].IEEE/ACM Transactions on Audio Speech and Language Processing,2025,33:3202–3214.
[30] LU Y J,WANG Z Q,WATANABE S,et al.Conditional diffusion probabilistic model for speech enhancement[C]//IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP).Singapore:IEEE,2022,7402–7406.
[31] RICHTER J,DE OLIVEIRA D,GERKMANN T.Investigating training objectives for generative speech enhancement[C]//IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP).Hyderabad,India:IEEE,2025,1–5.
[32] WANG T,CHEN R,LEI T,et al.Target matching based generative model for speech enhancement[J].IEEE/ACM Transactions on Audio Speech and Language Processing,2026,34:1633–1647.
[33] LU Y X,AI Y,LING Z H.Explicit estimation of magnitude and phase spectra in parallel for high-quality speech enhancement[J].Neural Networks,2025,189:107562.