Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2026, 15(2), 581-592
Published Online March 2026 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2026.152056

Hans汉斯

基于U-Net网络的MVDR声源定位伪峰抑制方法

蒋钦宇，李红莲，肖　瑶*，任志文，武欣艺

北京信息科技大学信息与通信工程学院，北京

收稿日期：2026年2月26日；录用日期：2026年3月12日；发布日期：2026年3月24日

摘　要

室内混响环境下的声源定位一直是声学信号处理领域的难点问题。传统的最小方差无失真响应(MVDR)波束形成算法在计算空间功率谱时，由于多径反射的影响，常在非声源位置产生大量伪峰，导致真实声源被淹没或混淆，严重制约了复杂环境下的声源识别能力。本文提出一种基于U-Net深度学习网络的MVDR空间谱后处理方法，将伪峰抑制问题转化为图像去噪任务。该方法以含混响的MVDR空间谱为输入，通过改进的U-Net网络学习从观测谱恢复理想谱的映射关系。网络引入残差结构、空间注意力机制和噪声抑制模块，并设计了结合全局重建、声源增强与伪峰抑制的复合损失函数。仿真实验表明，该方法能精准剥离混响引发的虚假伪影，在保持真实声源结构完整性的同时极大降低了误检风险，从而显著提升了复杂声场中声源目标的判别能力与定位鲁棒性。

关键词

声源定位，最小方差无失真响应，U-Net，深度学习，混响抑制

Pseudo-Peak Suppression for MVDR Sound Source Localization Based on U-Net

Qinyu Jiang, Honglian Li, Yao Xiao*, Zhiwen Ren, Xinyi Wu

School of Information and Communication Engineering, Beijing Information Science & Technology University, Beijing

Received: February 26, 2026; accepted: March 12, 2026; published: March 24, 2026

*通讯作者。

文章引用: 蒋钦宇, 李红莲, 肖瑶, 任志文, 武欣艺. 基于 U-Net 网络的 MVDR 声源定位伪峰抑制方法[J]. 人工智能与机器人研究, 2026, 15(2): 581-592. DOI: 10.12677/airr.2026.152056

Abstract

Sound source localization in indoor reverberant environments remains a critical challenge in the field of acoustic signal processing. When calculating spatial power spectra, the traditional Minimum Variance Distortionless Response (MVDR) beamforming algorithm often generates numerous pseudo-peaks at non-source locations due to multipath reflections. Consequently, true sound sources are frequently masked or obfuscated, severely compromising source identification capabilities in complex acoustic environments. To address this, this paper proposes an MVDR spatial spectrum post-processing method based on a U-Net deep learning network, formulating the pseudo-peak suppression problem as an image denoising task. Taking the reverberant MVDR spatial spectrum as input, the method employs an improved U-Net to learn the mapping relationship required to recover the ideal spectrum from observed data. The network incorporates residual blocks, a Spatial Attention Mechanism, and a Noise Suppression Module. Furthermore, a composite loss function is designed to synergize global reconstruction, source enhancement, and pseudo-peak suppression. Simulation results demonstrate that the proposed method accurately strips away reverberation-induced artifacts and preserves the structural integrity of the true source. By substantially reducing the risk of false detections, the method significantly enhances both the identifiability of sound sources and the robustness of localization in complex sound fields.

Keywords

Sound Source Localization, Minimum Variance Distortionless Response, U-Net, Deep Learning, Reverberation Suppression

1. 引言

声源定位技术在机器人听觉、人机交互及工业诊断等领域具有广泛的应用价值。它利用麦克风阵列估计声源的空间位置，但在室内等复杂声学环境中，多径传播导致的混响效应会严重干扰直达声，影响声源的判别并降低定位精度。

波束形成是声源定位的经典方法之一，其中最小方差无失真响应(Minimum Variance Distortionless Response, MVDR)算法因其高分辨率和强旁瓣抑制能力而备受关注[1] [2]。然而，在强混响条件下，MVDR算法生成的空间功率谱中常常会因相干多径反射而产生大量伪峰。这些伪峰并非真实声源，却会严重误导后续的峰值检测环节，导致定位结果出现较大偏差。

为抑制空间谱中的伪峰，研究者们提出了多种处理方法。一类是基于反卷积的传统信号处理技术，如 CLEAN 和 DAMAS 及其衍生算法[3]-[6]。这类方法将观测的声压图建模为真实声源分布与点扩散函数的卷积，通过迭代求解来重构一个更清晰的声源图像。尽管它们在一定程度上能提升分辨率并减少伪峰，但通常计算复杂度高，且其性能依赖于对声场模型的精确假设，在混响环境下稳健性不足。

另一条技术路径是结合深度学习方法。近年来，深度神经网络被广泛用于声学信号处理，展现出强大的非线性建模能力。在声源定位任务中，研究主要集中于两个方向：一是直接从多通道音频特征中学习端到端的定位模型，以应对混响和噪声[7] [8]；二是通过神经网络估计时频掩蔽或空间协方差矩阵，以

增强传统波束形成算法的性能[9] [10]。此外，U-Net 网络在处理声学映射图及语音增强任务中展现出显著的空间细节保持能力，为声学图像的后处理提供了新思路[11] [12]。这些方法验证了深度学习在处理复杂声学场景中的潜力，但大多致力于特征增强或端到端映射，较少专门针对 MVDR 等算法产生的空间谱中的伪峰进行抑制。

针对上述问题，本文提出一种谱生成–学习抑制处理框架。我们引入一个基于 U-Net 架构的深度神经网络作为后处理模块，将 MVDR 空间谱伪峰抑制问题转化为一个图像处理任务。该方法的核心思想是：利用 U-Net 强大的上下文信息提取与空间细节保持能力，学习从带有伪峰的观测谱到纯净参考谱之间的非线性映射关系。通过这种方式，网络能够有效识别并抑制由混响引起的虚假峰值，同时保持真实声源峰的锐度和位置精度。

2. 基于 U-Net 的声源定位方法

本章提出的声源定位系统整体框架如图 1 所示。该框架将含混响的 MVDR 空间功率谱作为输入特征，利用改进的 U-Net 网络进行端到端的非线性映射，以输出理想空间谱。

系统流程包含三个阶段。首先在数据生成阶段，在房间声学仿真的环境下，用镜像源法模拟混响产生，最终通过声源定位算法生成含混响噪声的观测谱与无干扰的理想谱，一一对应，形成数据集；网络训练阶段，以理想空间谱为监督目标，通过最小化损失函数优化网络参数，赋予模型特征增强与伪影抑制能力；推断定位阶段，利用未参与训练的仿真验证集数据，输入预训练模型进行前向推理。通过对比输出谱与理想谱，评估模型的伪峰抑制能力。

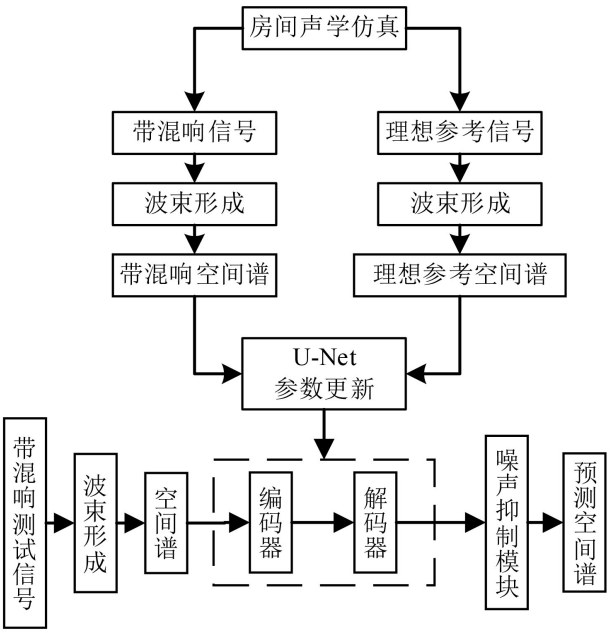

Figure 1. Overall workflow of the U-Net-based sound source localization algorithm
图 1. 基于 U-Net 的声源定位算法总流程

2.1. 基于 MVDR 的空间谱计算

首先将整个搜索空间划分为离散的网格点，每一个网格点都代表一个可能的声源位置。MVDR 算法通过在这些网格点上逐点扫描，计算其对应的空间功率值，从而生成完整的空间谱分布图。

假设麦克风阵列共有 K 个阵元，第 k 个阵元接收到的信号记为 $x_k(t)$，那么阵列接收到的信号为：

$$\boldsymbol{x}(t)=\left[x_1(t),x_2(t),\cdots,x_K(t)\right]^{\mathrm{T}} \tag{1}$$

对其进行短时傅里叶变换，得到时频域信号 $\boldsymbol{X}(f,t)$。对于搜索空间中的某一候选位置 $\boldsymbol{r}=[x,y,z]^{\mathrm{T}}$，设一共有 M 个麦克风，第 m 个麦克风的位置为 $\boldsymbol{r}_m=[x_m,y_m,z_m]^{\mathrm{T}}$，它的导向向量为：

$$\boldsymbol{a}(f,\boldsymbol{r})=\left[\frac{1}{d_1(\boldsymbol{r})}\mathrm{e}^{-j2\pi f\tau_1(\boldsymbol{r})},\frac{1}{d_2(\boldsymbol{r})}\mathrm{e}^{-j2\pi f\tau_2(\boldsymbol{r})},\cdots,\frac{1}{d_M(\boldsymbol{r})}\mathrm{e}^{-j2\pi f\tau_M(\boldsymbol{r})}\right]^{\mathrm{T}} \tag{2}$$

其中 $\tau_m(\boldsymbol{r})$ 表示声源到第 m 个麦克风的传播时延，$d_m(\boldsymbol{r})=\|\boldsymbol{r}-\boldsymbol{r}_m\|_2$，表示声源到第 m 个麦克风的实际空间几何距离。接着，根据接收信号估计空间协方差矩阵：

$$\boldsymbol{R}_x(f)=E\left[\boldsymbol{X}(f,t)\boldsymbol{X}^H(f,t)\right] \tag{3}$$

MVDR 算法的思想是寻找一个最优波束形成权向量 $\boldsymbol{\omega}_{\mathrm{MVDR}}$，在保持目标方向信号无失真通过的前提下，最大程度地抑制其他方向的干扰和噪声能量。该约束优化问题对应的最优权向量闭式解为：

$$\boldsymbol{\omega}_{\mathrm{MVDR}}(f,\boldsymbol{r})=\frac{\boldsymbol{R}_x^{-1}(f)\boldsymbol{a}(f,\boldsymbol{r})}{\boldsymbol{a}^H(f,\boldsymbol{r})\boldsymbol{R}_x^{-1}(f)\boldsymbol{a}(f,\boldsymbol{r})} \tag{4}$$

从而可得到该位置的 MVDR 空间功率谱值：

$$P(f,\boldsymbol{r})=\frac{1}{\boldsymbol{a}^H(f,\boldsymbol{r})\boldsymbol{R}_x^{-1}(f)\boldsymbol{a}(f,\boldsymbol{r})} \tag{5}$$

为构建适用于神经网络输入的图像特征，需对目标观测区域进行离散化处理。设被扫描区域为二维平面，沿 x 轴、y 轴分别均匀划分 N_x、N_y 个网格点，总采样点数为 $N=N_x\times N_y$，记第 i 行、第 j 列个网格点坐标为 $\boldsymbol{r}_{ij}$。该点处的 MVDR 功率谱值按式(5)计算：

$$S_{ij}=P(f,\boldsymbol{r}_{ij}) \tag{6}$$

式中 S_{ij} 表示网格点 (i,j) 处的声源能量强度；$i\subset[1,N_x]$；$j\subset[1,N_y]$。

遍历整个搜索平面，将 S_{ij} 按其空间坐标排列，形成 MVDR 空间谱矩阵：

$$\boldsymbol{S}=\begin{bmatrix}S_{11}&\cdots&S_{1N_y}\\ \vdots&\ddots&\vdots\\ S_{N_x1}&\cdots&S_{N_xN_y}\end{bmatrix}\in\mathbb{R}^{N_x\times N_y} \tag{7}$$

在该矩阵中，每个元素数值的大小反映了阵列在对应空间位置接收到的声源能量强弱。在理想情况下，真实声源位置对应的谱值将形成明显的主峰，从而构成一幅反映声场能量分布的二维声学图像[13]。

2.2. 改进的 U-Net 网络结构

本文设计的改进 U-Net 网络旨在捕捉真实声源与混响伪峰在空间能量分布上的结构性差异。为适配网络输入，首先将空间谱矩阵 $\boldsymbol{S}$ 进行归一化处理，得到网络的输入特征张量。

$$\boldsymbol{I}_{\mathrm{in}}=N(\boldsymbol{S}) \tag{8}$$

其中 $\boldsymbol{I}_{\mathrm{in}}$ 表示输入网络的单通道图像特征；$N(\cdot)$ 表示归一化操作。

网络整体基于经典的编码器–解码器架构，如图 2 所示。

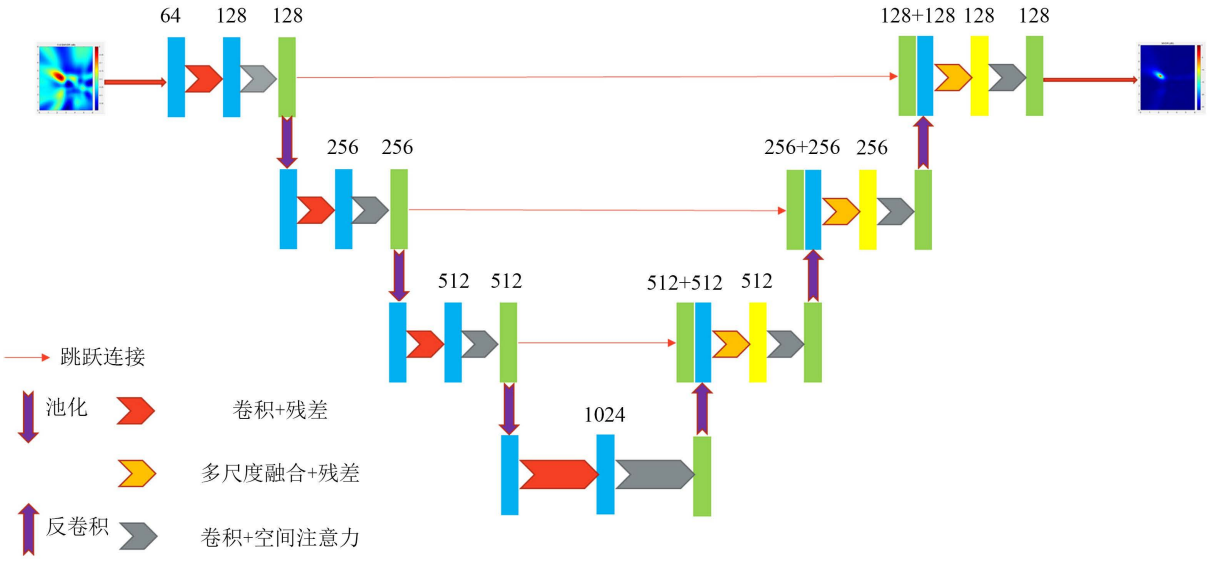

Figure 2. U-Net network architecture
图 2. U-Net 网络结构

2.2.1. 编码器设计

编码器负责提取空间谱的深层特征。主要由初始卷积块与 3 个级联的残差注意力下采样模块构成(如图 3 所示)。

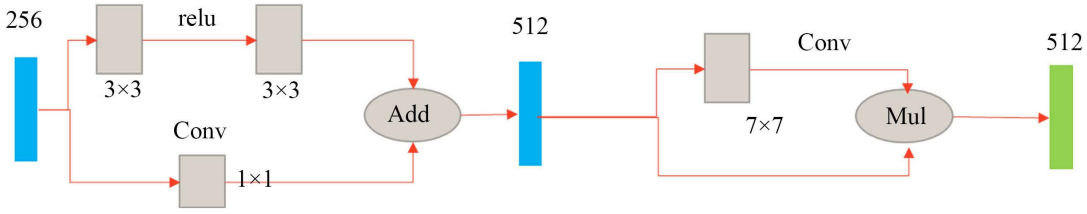

Figure 3. Encoder structure
图 3. 编码器结构

为解决深层网络训练中的梯度消失与特征退化问题，本文引入 ResNet 思想[14]，用残差模块替代经典 U-Net 的普通卷积单元。改进后的残差单元包含两个串联的 3 × 3 卷积层，并通过恒等映射实现特征的跳跃连接。记输入特征为 $\boldsymbol{F}_{\text{in}}$，则残差输出特征为 $\boldsymbol{F}_{\text{res}}$ 的计算过程为：

$$\boldsymbol{F}_{\text{res}} = \sigma\left(BN\left(\boldsymbol{W}_{\text{ext}} * \sigma\left(BN\left(\boldsymbol{W}_{\text{ext}} * \boldsymbol{F}_{\text{in}}\right)\right)\right)\right) + \boldsymbol{F}_{\text{in}} \tag{9}$$

其中，$*$ 表示卷积操作，$\sigma(\cdot)$ 为激活函数，$BN(\cdot)$ 为批归一化操作，$\boldsymbol{W}_{\text{ext}}$ 为 3 × 3 卷积核权重。为了方便，记为 $\boldsymbol{F}_{\text{res}} = H_{\text{res}}\left(\boldsymbol{F}_{\text{in}}\right)$，$H_{\text{res}}$ 代表式(9)定义的残差操作。

为缓解连续下采样造成的空间定位信息丢失，每个残差单元后均嵌入空间注意力模块。针对真实声源与伪峰的能量分布差异，本文采用 7 × 7 卷积核生成单通道空间权重图。相比于常规的小卷积核，大感受野能够完整覆盖声源主瓣区域，赋予网络由局部形态判别全局结构的能力。计算过程如下：

$$\boldsymbol{M}_{\text{att}} = \sigma_{\text{s}}\left(W_{\text{s}} * \boldsymbol{F}_{\text{res}}\right) \tag{10}$$

$$\boldsymbol{F}_{\text{att}} = \boldsymbol{M}_{\text{att}} \odot \boldsymbol{F}_{\text{res}} \tag{11}$$

其中，W_s 为 7×7 的卷积核权重；$\sigma_s(\cdot)$ 为 Sigmoid 激活函数，用于将权重映射至 $[0,1]$ 区间 $\boldsymbol{M}_{\text{att}}$ 为生成的空间注意力掩膜；$\odot$ 表示哈达玛积(Hadamard Product，即元素对元素相乘)；$\boldsymbol{F}_{\text{att}}$ 为经过注意力加权后的特征图。该机制在特征编码阶段可以对非声源区域的背景噪声进行抑制。

特征图经过注意力加权后，通过 2×2 最大池化层进行下采样。当空间谱经过三级编码模块以后，原始 128×128 的空间谱被压缩至 16×16 的瓶颈层特征，为后续解码过程提供了高度抽象的语义表征，各层特征尺度变化详见表 1。

Table 1. Scale changes in the encoding stage
表 1. 编码环节尺度变化

阶段	空间尺寸($H\times W$)	通道数 C
输入层	128 × 128	64
编码器第 1 层输出	64 × 64	128
编码器第 2 层输出	32 × 32	256
编码器第 3 层输出	16 × 16	512

2.2.2. 解码器设计

网络的瓶颈层采用了残差注意力，将特征通道数扩展至 1024，整合全局上下文信息来辅助真实声源与虚假声源的判别。解码器则通过 3 个上采样模块逐步恢复空间分辨率。首先，利用转置卷积将特征图尺寸放大，并与编码器对应层的浅层特征在通道维度进行拼接，来弥补下采样丢失的细节。

为了增强特征重用的效率，本文在解码块中设计了多尺度融合机制。假设解码器当前层的上采样拼接特征为 $\boldsymbol{F}_{\text{cat}}$，该机制包含两条路径：主路径通过残差块提取深层语义，旁路通过 1×1 卷积进行特征投影。两者相加实现多尺度信息的整合，计算公式如下：

$$\boldsymbol{F}_{\text{fus}} = H_{\text{res}}\left(\boldsymbol{F}_{\text{cat}}\right) + \boldsymbol{W}_{\text{proj}} * \boldsymbol{F}_{\text{cat}} \tag{12}$$

其中，$\boldsymbol{F}_{\text{fus}}$ 为融合后的特征张量；$\boldsymbol{W}_{\text{proj}}$ 为 1×1 卷积核权重，用于调整通道维度，从而匹配残差的输出。融合后的特征随后输入到空间注意力模块，利用空间注意力机制对高能量声源区域进行精准聚焦，确保网络在恢复空间分辨率的过程中，能够还原声源的主瓣能量分布并有效抑制背景干扰。整体结构如图 4 所示。

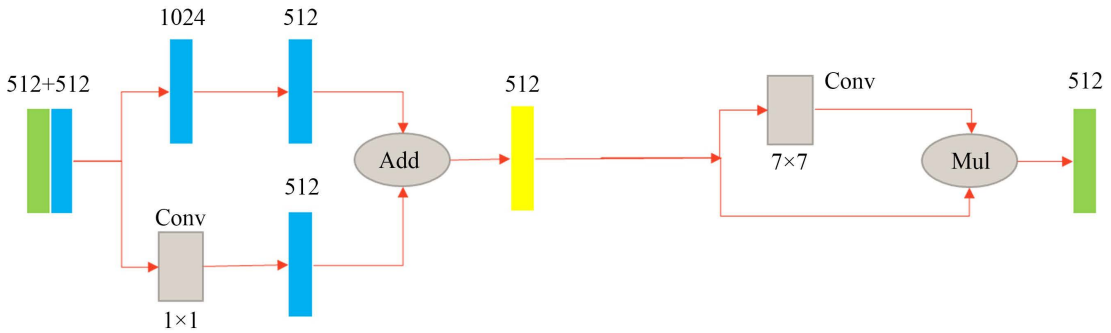

Figure 4. Decoder structure
图 4. 解码器结构

2.2.3. 噪声抑制模块

在网络的最终输出端，本文设计了一个噪声抑制模块(Noise suppression module, NSM)。首先，它利用 3 × 3 卷积提取局部噪声特征，经非线性激活后，再通过 1 × 1 卷积生成单通道的噪声概率图 $\boldsymbol{M}_{\text{noise}}$。最终，通过反向掩膜操作获得纯净的空间谱 $\boldsymbol{I}_{\text{out}}$。整体计算过程如下：

$$\boldsymbol{M}_{\text{noise}} = \sigma_{\text{s}}\left(\boldsymbol{W}_{\text{proj}} * \sigma\left(BN\left(\boldsymbol{W}_{\text{ext}} * \boldsymbol{F}_{\text{dec}}\right)\right)\right) \tag{13}$$

$$\boldsymbol{I}_{\text{out}} = \boldsymbol{F}_{\text{dec}} \odot \left(\boldsymbol{1} - \boldsymbol{M}_{\text{noise}}\right) \tag{14}$$

其中 $\boldsymbol{1}$ 表示全 1 矩阵。

2.3. 复合损失函数设计

在神经网络的训练中，本文设计了结合声源定位背景的损失函数，公式如下：

$$L = \lambda_{\text{global}} L_{\text{global}} + \lambda_{\text{src}} L_{src} + \lambda_{\text{art}} L_{\text{art}} \tag{15}$$

其中 λ 为各分项权重。首先，全局重建损失 L_{global} 的目的是对整张空间谱进行基础约束，它不仅利用 L_1 范数的稀疏诱导性加速非真实声源区域趋近于零，而且依靠均方误差确保真实声源主峰被保留。同时，为了避免声源主瓣成像模糊，将优化范围锁定在以声源为中心的 21×21 邻域内，通过对局部区域进行误差计算，引导网络精准重构声源的主瓣结构与峰值强度。此外，针对非声源区域的混响干扰，伪影抑制损失 L_{art} 利用反向空间掩膜，定向抑制远离声源中心的背景噪点。

3. 仿真设计与评价体系

本章通过仿真实验验证改进的 U-Net 网络在混响环境下的伪峰抑制效果。首先介绍基于镜像声源法的数据集构建与网络训练设置，随后通过可视化对比、定量指标及消融实验对模型性能进行全面评估。

3.1. 仿真环境与数据集构建

为了模拟真实的室内声传播环境，本文利用 RIR-Generator 工具包[15]生成多通道房间冲激响应，再将纯净语音信号与冲激响应卷积以合成麦克风阵列的接收信号，最后基于该信号计算 MVDR 空间谱作为网络输入。仿真房间尺寸设定为 8 m × 6 m × 5 m。接收端采用一个由 16 个全向麦克风组成的十字形阵列，阵元间距 0.05 m、孔径 0.35 m，阵列中心被固定在坐标(4 m, 3 m, 2.5 m)。信号采样率设定为 24 kHz。MVDR 算法的扫描平面固定在距离阵列中心 1.0 m 处的近场区域，相对区域范围为$[-2, 2] \times [-2, 2]$ m，混响时间设置为 0.3 s，共生成 3200 组原始样本。

预处理阶段，鉴于 Jet 色图红通道(R-Channel)承载主要能量且背景值近 0，仅提取 R 通道作为单通道输入。数据处理流程如下：将像素值线性归一化至$[0, 1]$，利用双线性插值统一调整尺寸至 128×128。最终，数据集按照 8:2 的比例划分为训练集(2560 组)和验证集(640 组)。

3.2. 网络训练设置

实验基于 PyTorch 2.3.0 框架，硬件搭载 AMD Ryzen 7 7840H 处理器与 NVIDIA RTX 4060 显卡。在训练过程中，采用 Adam 优化器进行网络参数更新，初始学习率设为 0.001，权重衰减设为1×10^{-5} 防止过拟合。训练批次大小设为 8，共进行 60 个 Epoch 的迭代训练。针对本文提出的复合损失函数，依据经验及多次调试，将各分项权重分别设置为：全局重建权重 $\lambda_{\text{global}} = 1.0$，声源保真权重 $\lambda_{\text{src}} = 0.5$，以及伪影抑制权重 $\lambda_{\text{art}} = 0.5$。

3.3. 评价指标

为了全面量化评估模型在混响环境下对 MVDR 空间谱的重构质量与伪峰抑制能力，本文采用了多项评价指标，分别从图像视觉保真度和声学信号特性两个维度进行考量。

首先，采用峰值信噪比(Peak Signal-to-Noise Ratio, PSNR)与结构相似性(Structural Similarity, SSIM)来衡量输出空间谱与理想无混响谱之间的图像级差异。PSNR 通过计算最大像素强度与均方误差的比值来反映重构图像的纯净度，其计算公式为：

$$R_{\mathrm{PSNR}} = 10 \cdot \log_{10}\left(\frac{I_{\mathrm{MAX}}^2}{R_{\mathrm{MSE}}}\right) \tag{16}$$

其中，I_{MAX} 为图像像素的最大可能值，R_{MSE} 为预测谱与真实谱之间的均方误差。SSIM 则从亮度、对比度和结构三个角度综合评价图像的感知质量[16]，数值越接近 1 表示网络输出在结构纹理上越接近理想声源分布。

其次，针对声源定位任务，本文定义了信噪比改善度(Signal-to-Noise Ratio Improvement, SNRI)与伪峰抑制率(Artifact Inhibition Ratio, AIR)两项声学指标。计算时构建背景掩码 M_{bg}，将距声源中心 6 像素以外区域定义为背景干扰区。

SNRI 用于评估输出空间谱的动态范围，即主峰能量相对于背景噪声水平的显著程度，其计算如下：

$$R_{\mathrm{SNRI}} = 20 \cdot \log_{10}\left(\frac{\max\left(P_{\mathrm{out}}\right)}{\mu_{\mathrm{bg}}}\right) \tag{17}$$

其中 $\max\left(P_{\mathrm{out}}\right)$ 为输出谱的峰值强度，μ_{bg} 为背景掩码区域内像素的均值。SNRI 的值越高，表明声源主峰越突出，背景越干净。

AIR 则专门用于量化模型对伪峰的抑制能力，通过计算背景区域能量在处理前后的变化率得出：

$$R_{\mathrm{AIR}} = \max\left(0, 1 - \frac{E_{\mathrm{out}}^{\mathrm{bg}}}{E_{\mathrm{in}}^{\mathrm{bg}}}\right) \times 100\% \tag{18}$$

$E_{\mathrm{in}}^{\mathrm{bg}}$ 和 $E_{\mathrm{out}}^{\mathrm{bg}}$ 分别代表输入谱和输出谱在背景区域内的能量总和。AIR 的值越高，意味着模型去除伪峰效果越好。

此外，为了直接量化该方法在声源定位精度上的实际表现，本文引入均方根误差(Root Mean Square Error, RMSE)作为绝对定位评价指标。对于空间功率谱，提取全局能量最大峰值所对应的物理空间坐标作为预测定位点，RMSE 的计算公式如下：

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\left(\hat{x}_i - x_i\right)^2 + \left(\hat{y}_i - y_i\right)^2\right]} \tag{19}$$

其中，N 为测试集样本总数；$\left(x_i, y_i\right)$ 为真实声源的坐标；$\left(\hat{x}_i, \hat{y}_i\right)$ 为模型输出空间谱中最高峰对应的预测坐标。RMSE 值越小，表明算法的定位偏差越小、精度越高。

4. 实验结果与分析

4.1. 空间谱可视化分析

为了直观地评估 U-Net 模型在混响环境下对声源定位的优化性能，本节对模型处理前后的空间谱进行了可视化对比分析，结果如图 5 所示。

图 5 的第一行展示了受混响与噪声共同干扰的原始空间谱。可以清晰地观察到，复杂的声学环境导

致了严重的成像质量下降，具体表现为：1) 出现了多个偏离真实声源位置的伪峰；2) 真实声源所对应的主峰发生了显著的定位偏移；3) 主峰能量弥散，呈现出明显的展宽现象，降低了空间分辨率。

经过本模型处理后，如第二行所示，伪峰被有效抑制，空间谱质量得到了显著提升。同时，模型成功地对主峰的位置和形态进行了恢复，不仅修正了定位偏移，而且重塑了能量集中的主峰轮廓。该结果证明，本文所提出的方法能够从含伪峰的空间谱中恢复出清晰、无伪峰的空间谱，增强了声源定位算法的鲁棒性与可靠性。

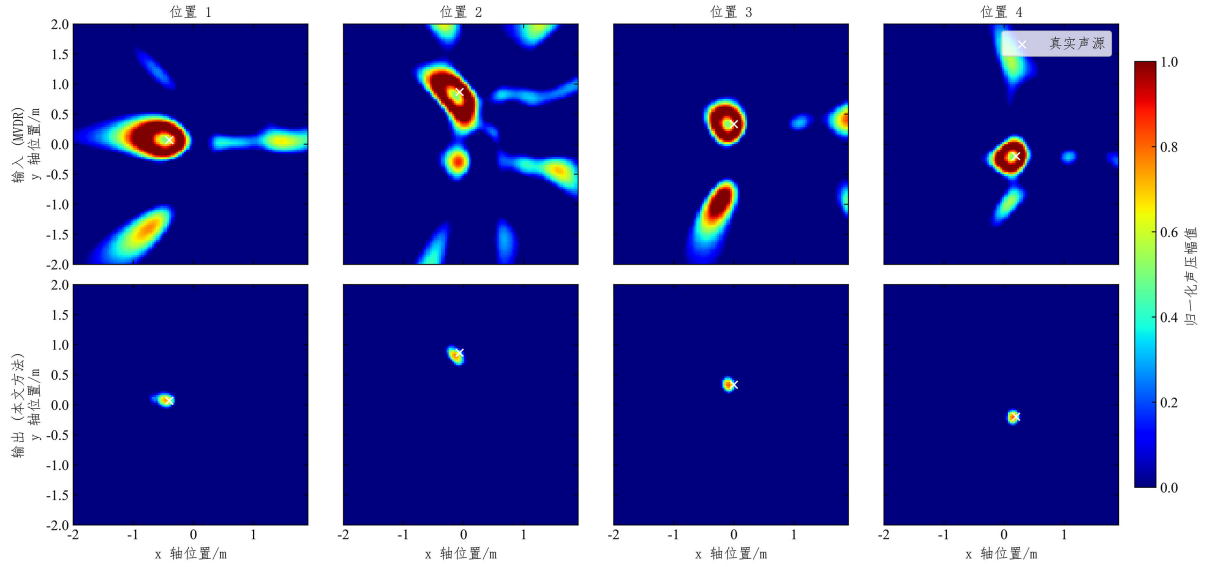

Figure 5. Comparison of U-Net post-processing results on the MVDR spatial spectrum
图 5. U-Net 网络对 MVDR 空间谱的后处理结果对比

4.2. 定量评估指标

为了从统计角度进一步验证本文所提方法在混响环境下的综合性能，本节选取了峰值信噪比(PSNR)、结构相似度(SSIM)、伪峰抑制率(AIR)、信噪比改善度(SNRI)以及均方根误差(RMSE)五个指标，对测试集上所有验证样本的前向推理结果进行了统计与定量评估。本文方法在各项评价指标上的平均测试结果如表 2 所示。

Table 2. Performance metrics
表 2. 性能指标

评估指标	测量值
PSNR	30.47 dB
SSIM	0.9753
AIR	81.80%
SNRI	36.64 dB
RMSE	0.05 m

基于表 2 的统计数据可以得出，本文提出的基于改进 U-Net 的空间谱后处理方法在各个维度上均展现出了优异的性能。在图像重建质量方面，模型的平均 PSNR 达到了 30.47 dB，同时 SSIM 高达 0.9753，

这说明网络不仅有效去除了图像噪点，更极好地保持了声源主瓣的形态结构，未出现过度平滑或结构丢失现象。

在声学特性与干扰抑制方面，得益于网络末端设计的噪声抑制模块(NSM)对非声源区域的强力过滤，模型的伪峰抑制率(AIR)达到了 81.80%，意味着输入原始谱中绝大多数的伪峰已被网络成功剥离。与此同时，SNRI 指标达到了 36.64 dB，输出空间谱的动态范围得到了极大扩展。这种高对比度输出使得声源主峰从背景中高度凸显，从根本上降低了后续寻峰算法的误检率。

最后，在绝对定位精度方面，本模型在 128 × 128 离散网格划分(物理范围 4 m × 4 m)下，将全局最大峰值对应的空间几何误差(RMSE)严格控制在了 0.05 m 左右。该指标直接证明了本文算法在强混响干扰下，依然具备极高且稳健的声源坐标锁定能力。

4.3. 消融实验

为了验证噪声抑制模块和复合损失函数设计的必要性，本节对比了完整模型、移除噪声抑制模块、以及仅使用全局损失函数三种配置下的性能表现，结果如图 6 所示。

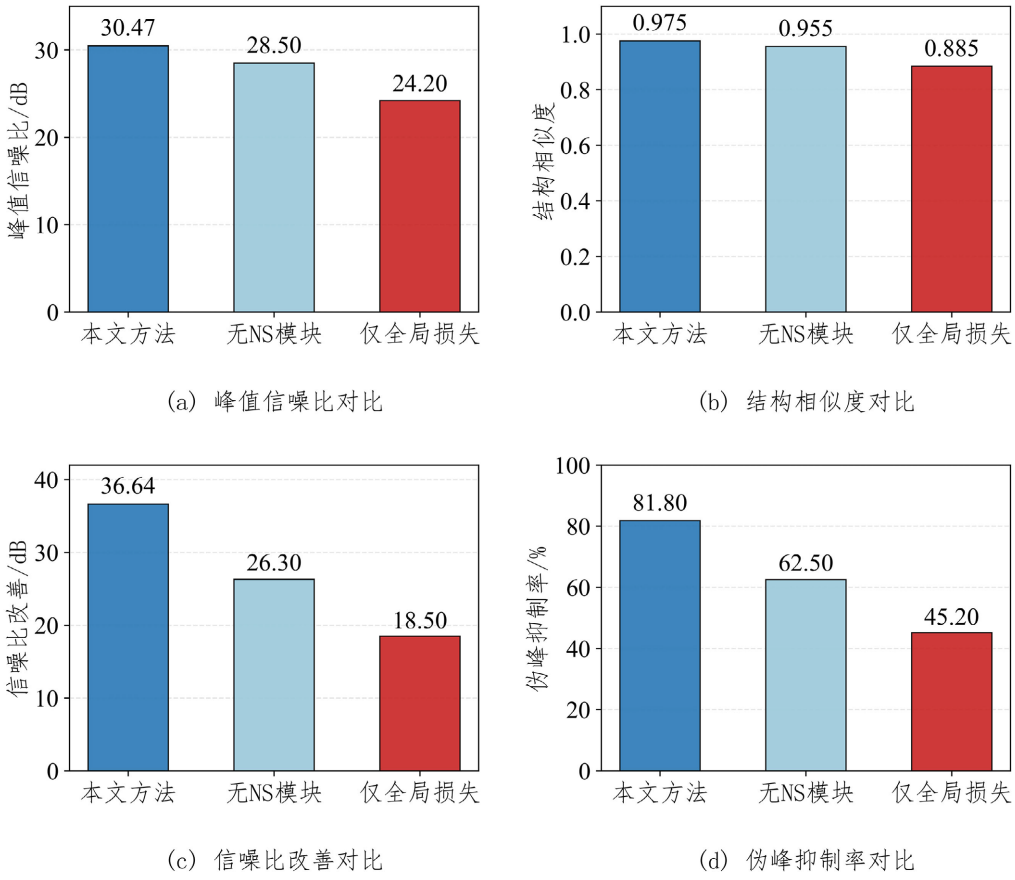

Figure 6. Comparison of ablation experiment results
图 6. 消融实验结果对比

实验结果明确了各模块的贡献。首先，移除 NSM 模块导致性能显著下降：AIR 从 81.80%跌至 62.50%，SNRI 由 36.64 dB 降至 26.30 dB。这证实了 NSM 模块通过反向掩膜机制，在抑制伪峰方面起到了决定性作用。

另一方面，仅依赖全局损失函数的模型表现最差。单一的全局约束难以兼顾局部细节，导致声源主瓣模糊且空间谱存在背景噪声残留。相比之下，复合损失函数通过引入局部保真与背景抑制项，在确保高图像信噪比的同时，有效重构了空间谱。综上，NSM 模块与复合损失函数的协同作用是本方法性能优势的基础。

5. 结论

针对室内混响环境下 MVDR 算法易生伪峰的问题，本文提出一种基于改进 U-Net 的后处理方法。研究表明，该方法有效建立了观测谱与理想谱间的非线性映射，在保持高结构相似度(SSIM 0.9753)的同时实现了 81.80%的伪峰抑制率，验证了将声源定位修正转化为图像去噪任务的有效性。

消融实验证实，噪声抑制模块(NSM)与复合损失函数是保障性能的核心。移除 NSM 导致伪峰抑制能力显著下降，而单一全局损失无法兼顾声源细节与背景纯净度。这表明针对声场特性的专用网络结构在剥离混响伪影、重塑峰值形态方面具有关键作用，确保了系统的高信噪比与鲁棒性。

目前研究主要局限于单声源仿真场景。未来工作将拓展多声源检测能力，并结合迁移学习，验证模型在真实复杂环境及嵌入式平台上的泛化性能与实时性，以推动更广泛的工程应用。

参考文献

[1] Capon, J. (1969) High-Resolution Frequency-Wavenumber Spectrum Analysis. *Proceedings of the IEEE*, **57**, 1408-1418. https://doi.org/10.1109/proc.1969.7278

[2] Van Trees, H.L. (2002) Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory. Wiley. https://doi.org/10.1002/0471221104

[3] Wang, Y., Deng, Z., Zhao, J., Kopiev, V.F., Gao, D. and Chen, W. (2025) Progress in Beamforming Acoustic Imaging Based on Phased Microphone Arrays: Algorithms and Applications. *Measurement*, **242**, Article ID: 116100. https://doi.org/10.1016/j.measurement.2024.116100

[4] Lobato, T. and Sotteck, R. (2024) Accelerating the CLEAN-SC and CMF Beamforming Deconvolution Methods Using Neural Grid Compression. In: *INTER-NOISE and NOISE-CON Congress and Conference Proceedings* (*INTER-NOISE* 2024), Institute of Noise Control Engineering, art00003.

[5] Yardibi, T., Li, J., Stoica, P. and Cattafesta, L.N. (2008) Sparsity Constrained Deconvolution Approaches for Acoustic Source Mapping. *The Journal of the Acoustical Society of America*, **123**, 2631-2642. https://doi.org/10.1121/1.2896754

[6] Ning, F., Jia, D., Hou, H., Meng, D., Hao, M. and Wei, J. (2025) A High-Resolution Sparse Coherent Sound Source Localization Approach with Improved Sparsity Constraint. *Mechanical Systems and Signal Processing*, **232**, Article ID: 112712. https://doi.org/10.1016/j.ymssp.2025.112712

[7] Grumiaux, P., Kitić, S., Girin, L. and Guérin, A. (2022) A Survey of Sound Source Localization with Deep Learning Methods. *The Journal of the Acoustical Society of America*, **152**, 107-151. https://doi.org/10.1121/10.0011809

[8] Shimada, K., Koyama, Y., Takahashi, S., Takahashi, N., Tsunoo, E. and Mitsufuji, Y. (2022) Multi-ACCDOA: Localizing and Detecting Overlapping Sounds from the Same Class with Auxiliary Duplicating Permutation Invariant Training. *ICASSP* 2022-2022 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Singapore, 23-27 May 2022, 316-320. https://doi.org/10.1109/icassp43922.2022.9746384

[9] Kim, M., Cheong, S. and Shin, J.W. (2023) DNN-Based Parameter Estimation for MVDR Beamforming and Post-Filtering. *INTERSPEECH* 2023, Dublin, 20-24 August 2023, 3879-3883. https://doi.org/10.21437/interspeech.2023-420

[10] Kim, H., Kang, K. and Shin, J.W. (2022) Factorized MVDR Deep Beamforming for Multi-Channel Speech Enhancement. *IEEE Signal Processing Letters*, **29**, 1898-1902. https://doi.org/10.1109/lsp.2022.3200581

[11] Ren, X., Zhang, X., Chen, L., Zheng, X., Zhang, C., Guo, L., *et al*. (2021) A Causal U-Net Based Neural Beamforming Network for Real-Time Multi-Channel Speech Enhancement. *Proceedings of INTERSPEECH* 2021, Brno, 30 August-3 September 2021, 1832-1836. https://doi.org/10.21437/interspeech.2021-1457

[12] Jia, H., Yang, F., Hu, X. and Yang, J. (2025) A Dual-Encoder U-Net Architecture with Prior Knowledge Embedding for Acoustic Source Mapping. *The Journal of the Acoustical Society of America*, **158**, 1767-1782. https://doi.org/10.1121/10.0039104

[13] Merino-Martínez, R., Sijtsma, P., Snellen, M., Ahlefeldt, T., Antoni, J., Bahr, C.J., *et al*. (2019) A Review of Acoustic

Imaging Methods Using Phased Microphone Arrays. *CEAS Aeronautical Journal*, **10**, 197-230. https://doi.org/10.1007/s13272-019-00383-4

[14] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/cvpr.2016.90

[15] Habets, E.A.P. (2006) Room Impulse Response Generator. Technische Universiteit Eindhoven, 1-24.

[16] Wang, Z., Bovik, A.C., Sheikh, H.R. and Simoncelli, E.P. (2004) Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, **13**, 600-612. https://doi.org/10.1109/tip.2003.819861