中图分类号：　　单位代号：10280
密　　级：　　学　　号：22721436

上海大学

专业硕士学位论文

SHANGHAI UNIVERSITY PROFESSIONAL MASTER'S DISSERTATION

题目	抗声源分离攻击的鲁棒音频水印技术研究

作　　者：史景辉
学科专业：电子信息
导　　师：吴汉舟
完成日期：2025 年 5 月

姓　　名：史景辉　　学号：22721436

论文题目：抗声源分离攻击的鲁棒音频水印技术研究

上海大学

本论文经答辩委员会全体委员审查，确认符合上海大学硕士学位论文质量要求。

答辩委员会签名：

主　　席：

委　　员：

导　　师：

答辩日期：　　年　　月　　日

姓　　名：史景辉　　学号：22721436

论文题目：抗声源分离攻击的鲁棒音频水印技术研究

上海大学学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师指导下，独立进行研究工作所取得的成果。除了文中特别加以标注和致谢的内容外，论文中不包含其他人已发表或撰写过的研究成果。参与同一工作的其他研究者对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：

日 期：　　年　　月　　日

上海大学学位论文使用授权说明

本人完全了解上海大学有关保留、使用学位论文的规定，即：学校有权保留论文及送交论文复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容。

（**保密论文在解密后应遵守此规定**）

学位论文作者签名：　　导师签名：

日 期：　　年　　月　　日　　日 期：　　年　　月　　日

上海大学工程硕士学位论文

抗声源分离攻击的鲁棒音频水印技术研究

作　　者：史景辉

学科专业：电子信息

导　　师：吴汉舟

上海大学通信与信息工程学院

二〇二五年五月

A Dissertation Submitted to Shanghai University for the Degree of Master of Engineering

Robust Audio Watermarking Against Source Separation Attacks

Candidate： **Jinghui Shi**

Major： **Electronic Information**

Supervisor： **Hanzhou Wu**

School of Communication and Information Engineering

Shanghai University

May, 2025

摘　要

随着互联网和多媒体技术的发展，数字音频在提升人机交互效率和日常生活便利性的同时，也面临着诸如非法盗用、二次合成或篡改等安全挑战。声源分离技术用于从混合语音信号中分离出单个声源，其广泛应用加剧了上述威胁。音频水印技术可在不影响音频感知质量的前提下嵌入秘密信息，达到版权保护或溯源的目的。然而，现有的音频水印研究主要针对电子攻击（如压缩、加噪、滤波等），对于声源分离攻击的研究相对较少，一定程度上限制了现有方法的适用性。在此背景下，本文从时域和频域两个角度出发，研究抗声源分离攻击的鲁棒音频水印方案，具体内容如下：

1）针对音频中的声源分离攻击，提出了一种基于可逆神经网络的鲁棒音频水印算法。该算法将可逆神经网络作为核心架构，在水印的嵌入阶段，对水印信号进行扩频编码，并结合冗余编码技术将水印信息嵌入音频信号的时域中；在水印提取阶段，利用网络的可逆映射特性，采用与嵌入网络参数一致的网络模型，通过线性映射完成水印提取。为抵抗声源分离攻击，本文对声源分离过程中的失真进行了细致分析，设计了相应的模拟失真层来匹配所提出的水印框架。此外，模型在训练过程中引入了低频损失函数，提高了水印的不可感知性。实验表明，该算法能够在保持音频听觉质量的前提下将水印嵌入多声源音频，并对声源分离以及噪声添加、滤波、重采样等常见攻击表现出良好的抵抗能力。

2）考虑到在时域嵌入水印鲁棒性较差，提出了一种基于生成对抗网络的频域鲁棒音频水印算法。该算法使用短时傅里叶变换将音频变换到频域，并利用调制网络对水印信号进行调制和维度对齐。在该框架中，生成器以音频频谱和水印信号作为输入，使生成的含水印音频具有高保真特性；解码器从含水印音频中准确恢复秘密信息；鉴别器用于引入对抗训练，引导含水印音频在听觉质量上逼近原始音频。此外，通过引入强度因子实现了水印在不可感知性与鲁棒性之间的平衡。实验表明，该算法在抵抗常见信号攻击方面较空域嵌入表现出更强的鲁棒性，并能在声源分离场景下具备更强的抗干扰能力，同时具有良好的不可感知性。

关键词：可逆神经网络，生成对抗网络，音频水印，版权保护，声源分离

ABSTRACT

With the rapid development of the Internet and multimedia technologies, digital audio has become a critical medium for enhancing human–computer interaction and daily convenience. However, it also faces significant security challenges, such as unauthorized usage, re-synthesis, and tampering. Source separation techniques, which aim to isolate individual sources from mixed speech signals, have been widely adopted in various audio processing applications, further aggravating these security risks. Audio watermarking technology, which embeds imperceptible information into audio signals, offers a promising solution for copyright protection and content tracing. Nevertheless, most existing audio watermarking approaches primarily focus on resisting conventional signal processing attacks (e.g., compression, noise addition, and filtering), while research on robustness against source separation attacks remains limited. To address this gap, this dissertation proposes two robust audio watermarking schemes from both time-domain and frequency-domain, designed to withstand source separation attacks. The main contributions are as follows:

1) To deal with source separation attacks in audios, this dissertation proposes a robust audio watermarking technology based on invertible neural network. The core of the proposed method is an invertible neural network that enables accurate recovery of embedded watermarks. During the embedding stage, the watermark is spread-spectrum encoded and redundantly embedded into the time-domain audio signal. During the extraction stage, the reversible mapping of the same network is used to recover the watermark without requiring access to the original audio. To improve resilience, the distortion introduced by source separation is carefully modeled and simulated using a distortion layer within the network. Additionally, a low-frequency loss function is incorporated during training stage to enhance watermark imperceptibility. Experimental results demonstrate that the proposed method effectively embeds watermarks into multi-source audio while maintaining perceptual quality, and shows strong resistance to source separation, noise addition, filtering, resampling, and other common signal attacks.

2) To address the limitations of time-domain embedding in terms of robustness, this dissertation proposes a frequency-domain audio watermarking scheme based on

generative adversarial networks. The audio is first transformed into the frequency domain using the Short-Time Fourier Transform. Then a modulation network is used to align the watermark signal with the audio spectral features. Within this framework, the generator takes both the audio spectrum and the watermark as input, producing high-fidelity watermarked audio; the decoder accurately extracts the watermark from the embedded audio; and the discriminator guides adversarial training to ensure that the watermarked audio remains perceptually similar to the original. An embedding strength factor is introduced to balance the trade-off between imperceptibility and robustness. Experimental results show that this method exhibits stronger robustness than time-domain schemes under common signal processing attacks, and maintains reliable watermark extraction even in the presence of source separation, while ensuring high imperceptibility.

Keywords: Invertible Neural Network, Generative Adversarial Networks, Audio Watermarking, Copyright Protection, Sounds Source Separation

目　录

第一章　绪论

1.1 研究背景与意义

随着数字音频技术的飞速发展以及流媒体平台的普及[1]，音频作为信息传播和文化传播的重要载体，在社交媒体、在线教育、智能语音交互等多个领域得到了广泛应用[2,3]。与此同时，音频内容的制作与传播成本不断降低，大量音频数据在网络中高速流通，随之带来的便是其被复制、篡改、滥用等一系列安全问题[4,5]。

为了有效保障音频内容的版权与完整性，音频水印技术[6]应运而生。该技术是一种将特定信息嵌入到音频信号中的信息隐藏技术，这些信息在音频播放过程中几乎无法被察觉，但在需要时可通过特定算法提取，用于验证内容的合法性或追溯内容的来源[7,8]。相比于传统的加密技术，音频水印具有信息与内容紧密绑定的特点，即便音频内容在网络中被传播、复制，水印仍随之保留并具备可识别性。此外，水印嵌入后不影响用户的正常听觉体验，使其能够“隐形”地保留在音频中，因而被广泛应用于音频的版权保护[9]。随着应用需求的不断拓展，音频水印算法在实际应用中面临着日益严峻的信道干扰与攻击威胁，这对其鲁棒性提出了更高的要求。现有的音频水印方法主要针对信号处理类攻击如压缩、加噪、滤波、重采样等进行设计。但随着近年来声源分离技术(Source Separation，SS)[10,11]的兴起，音频水印技术面临新的安全挑战。

声源分离技术，尤其是基于深度学习[12]的盲音频分离方法，能够从混合音频中提取出独立的说话人语音或声音源，这项技术在语音识别[13-15]、语音增强[16]、人机对话[17]等领域发挥了重要作用。然而，这类技术也为现有的音频水印算法带来了新的安全威胁，如图 1.1 所示。对于已嵌入水印的音频，一方面，有意的攻击者可能利用声源分离工具“剥离”音频中已有水印成分，从而削弱或破坏水印信息的提取与验证；另一方面，声源分离的过程也可能在无意中破坏音频信号的水印结构。例如在多说话人录音中，为了提取某一说话人语音而进行的声源分离处理，可能在不知情的情况下破坏水印嵌入结构，导致水印信息难以提取，进而

影响版权追踪与内容认证的有效性。这种“无意中的攻击”往往容易对水印的实施造成更大的威胁。

图 1.1 声源分离对音频水印的挑战

基于上述原因，研究抗声源分离攻击的鲁棒音频水印算法，对于提升音频水印技术在现实复杂环境下的可靠性具有重要意义。这类技术不仅要在面对传统信号处理攻击时保持稳定的性能，还需在经历声源分离操作后仍然能准确提取水印，确保水印信息的可用性与可靠性。该研究有助于拓展现有水印技术在语音安全、通信认证等领域的应用边界。

1.2 国内外研究概况

1.2.1 音频水印技术研究现状

数字水印是一种应对信息安全问题的有效手段[18]，被广泛应用于数字版权保护、广播监测、身份认证、内容溯源等领域[19]。相较于图像或视频，音频信号的嵌入容量较小，且由于人耳对噪声较敏感的特性[20]，音频水印在不可感知性和鲁棒性之间的权衡成为研究难题。

传统的音频水印方法分为基于时域和基于变换域两种。基于时域的方法是指直接对音频的采样值进行修改，如替换最低有效位(Least Significant Bit, LSB)[21]算法将要嵌入的水印信息编码为二进制比特，再从音频文件中提取出每个样本点，然后将每个样本点的最低有效位替换为水印信息中的相应比特。例如，如果

水印信息的比特位为 0，将采样点的最低有效位修改为 0，相反，如果水印信息的比特为 1，则将最低有效位改为 1。替换最低有效位的音频水印算法最早由 Bender 等人[22]提出。后来 Cvejic 等人[23]基于心理声学模型对 LSB 算法进行了改进，提出了一种结合心理声学模型的 LSB 改进算法，利用人耳的听觉掩蔽效应动态选择水印嵌入位置，该方法优先在高能量音频帧的最低有效位嵌入水印，从而在不影响听感的情况下提高水印容量。Gulve 等人[24]提出一种动态调整 LSB 嵌入深度的方法，根据音频信号的能量或过零率，在高能量段和低能量段分别使用不同的 LSB 算法进行嵌入并结合汉明码进行纠错。

时域回声隐藏[25]法是另一种基于时域的音频水印方法。在该类方法中，通过在音频信号中引入微弱的回声信号实现水印嵌入，回声的特性（如延迟、衰减、相位等）用于表示不同的水印比特。由于回声的影响通常较小且难以察觉，这种方法在水印的不可感知性方面具有较大的优势。Bender 等人[22]提出一种基于回声隐藏的音频水印算法，其具体做法是在原始信号中叠加短延迟的回声，通过调整回声的延迟时间和幅度编码水印信息。Erfani 等人[26]提出了一种基于内容的改进型时间扩展回声隐藏音频水印算法，将解码器处原始信号倒谱部分误差消除，从而提高解码器的检测率和水印的不可感知性。这类基于时域的水印方法在水印嵌入过程中不涉及矩阵变换和复杂编码，处理过程只需要遍历采样值进行嵌入或检测，因此水印的嵌入效率高且计算量小。但这种算法的缺点也非常明显，尤其是在面对实际复杂应用时性能较差。

基于变换域的音频水印方法是指将音频信号从时域转换到频域，在频谱特定区域进行水印嵌入。首先对音频进行频域变换，常见的变换方法有离散小波变换(Discrete Wavelet Transform, DWT)、离散余弦变换(Discrete Cosine Transform, DCT)、快速傅里叶变换(Fast Fourier Transform, FFT)、离散傅里叶变换(Discrete Fourier Transform, DFT)以及音频信号处理中常用的短时傅里叶变换(Short-Time Fourier Transform, STFT)等，然后直接修改变换域内的系数矩阵或频谱，最后再通过相应的反变换将数据恢复至时域进行存储和传输等操作。Wang 等人[27]提出一种基于 DWT 与 DCT 的盲音频水印算法，通过将同步码嵌入样本均值实现同步，同时利用频域掩蔽特性将水印嵌入 DCT 低频系数中，提升了水印的不可感知性和在同步攻击、裁剪、压缩等条件下的鲁棒性。Lei 等人[28]首先对音频信号

应用 DWT 提取低频近似子带特征，再结合 DCT 和奇异值分解(Singular Value Decomposition, SVD)获取嵌入载体，通过差分进化算法(Differential Evolution, DE)优化量化步长以嵌入水印，从而在保证音质的同时提升了算法的鲁棒性。Hu 等人[29]提出一种基于 FFT 的双模盲音频水印算法，可同时在音频中隐藏二值图像与彩色图像，该方法结合自适应幅度调制与相位调制两种机制，其中自适应幅度调制通过调制低频系数幅度来嵌入二值水印，相位调制通过调节中频系数的相位嵌入彩色图像，该方法在不影响音质的前提下，在应对裁剪、压缩以及去同步等多种攻击均表现出良好的鲁棒性。变换域算法将音频信号映射到频率域或多尺度空间中，能提取出稳定、低冗余的特征，可以将水印嵌入在对攻击不敏感的区域中，所以对常见的攻击有更强的抵抗能力。而且还可以将水印嵌入在人类听觉不敏感的区域，使得带水印的音频拥有更高的不可感知性。但频域嵌入需要进行变换，通常算法计算复杂度较高，计算量相较于时域水印算法更大，实时性较差。

与上述传统音频水印算法不同，基于深度学习的音频水印算法是近年来兴起的技术，利用生成对抗网络、自动编码器等深度学习网络进行水印的嵌入和提取。此类算法的优势在于自适应性，能够根据输入音频数据和水印的需求自动优化嵌入策略。基于深度学习的数字水印主流框架由三部分组成：用于嵌入水印的编码层、用于提取水印的解码层，以及模拟各种攻击的失真层。Li 等人[30]提出了一种双重嵌入水印算法，将定位信息与水印信息分别嵌入，以加快定位过程，并通过引入平衡模块保持可逆神经网络(Invertible Neural Networks, INN)的对称性，在不可感知性、容量、鲁棒性及定位效率等方面取得了良好性能。Kosta 等人[31]设计了一种联合训练的嵌入器与检测器结构，两者目标相反却通过联合优化协同训练，嵌入器在保证音质的同时尽可能嵌入信息，检测器则高效提取水印，同时通过引入对抗结构实现系统鲁棒性与隐蔽性之间的权衡。Roman 等人[32]针对人工智能生成语音检测问题，设计了一种水印标记与检测系统，通过联合训练水印生成器和检测器，在不损害音频感知质量的前提下实现样本级检测，并引入基于听觉掩蔽效应的感知损失函数，使得水印嵌入更符合人类听觉特性。Chen 等人[33]构建了一种融合可逆神经网络、短时傅里叶变换与滑动窗口机制的音频水印框架，显著提升了水印的隐蔽性与恢复精度。该方法通过滑动窗口与穷举检测策略，无需同步码即可自动定位水印位置，避免了传统方法在同步过程中可能遭受的攻

击。Liu 等人[34]提出了一种面向音频重录攻击(Audio Re-recording, AR)的鲁棒音频水印算法，对传统算法在实际重录场景中性能失效的问题进行了系统性建模与优化，该系统通过联合训练的编码器与解码器，实现端到端的水印嵌入与提取，并设计了可微分的失真层模拟重录过程中的关键变化，从而显著增强系统在真实重录环境下的鲁棒性。Liu 等人[35]提出了一种面向语音克隆攻击的鲁棒音频水印算法，该方法通过端到端联合训练的嵌入器与提取器，将水印嵌入频域并使用失真层模拟语音克隆中的典型变化，在多种语音合成工具下成功提取出水印。在这些方法中，端到端训练至关重要，它可以引入自定义的失真层，增强对各种攻击的鲁棒性。

综上所述，音频水印技术已从早期基于时域和频域的手工特征工程方法逐渐转向以深度神经网络为核心的数据驱动的嵌入和提取框架。尽管传统方法在实现简便性与理论构建方面具有一定优势，但普遍面临嵌入容量受限、鲁棒性不足以及难以适应复杂信号环境的瓶颈。而深度学习方法凭借其强大的特征表示与非线性建模能力，通过端到端训练机制，自适应学习音频内容的局部结构与全局模式，自动优化嵌入位置与策略，在水印的不可感知性、嵌入容量以及鲁棒性等方面表现出明显优势。此外，失真模拟层使得基于深度学习的水印系统在面对各种攻击时仍能保持良好的鲁棒性。因此，相比传统水印方法，基于深度学习的音频水印技术具有更广阔的应用前景与发展潜力。

1.2.2 声源分离技术研究现状

声源分离作为音频信号处理领域的重要研究方向，用于从混合音频信号中提取出各个独立的声源成分。该技术在语音识别、语音增强、智能语音交互以及音乐信息检索等领域具有广泛的应用价值。当前主流的声源分离方法可分为三类：基于混合信号的盲源分离算法、基于模型的声源分离算法以及基于深度学习的声源分离算法。

基于混合信号的盲源分离是最早发展的声源分离技术之一，其核心思想是在缺乏先验知识的情况下，仅利用观测信号实现源信号的分离。Hyvärinen 等人[36]提出的独立成分分析(Independent Component Analysis, ICA)是声源分离领域的奠基性方法，其基本假设是观测信号由一组统计独立的源信号线性混合而成。该方

法通常通过最大化非高斯性等准则估计混合矩阵，从而实现对源信号的解混分离。然而，ICA 方法仅适用于线性瞬时混合场景，对实际环境中的卷积混音问题处理效果有限。为克服这一局限，Lee 等人[37]提出了非负矩阵分解(Non-negative Matrix Factorization, NMF)方法，该方法将频谱矩阵分解为两个非负矩阵，通过迭代优化逼近原始信号。NMF 方法充分利用了音频信号的非负性先验，在音乐分离任务中表现出色。需要指出的是，这类分离算法通常需要满足信号源相互独立、混合信号具有高斯分布等严格假设，这在一定程度上限制了其在实际复杂场景中的应用效果。

基于模型的声源分离算法通过构建特定的声学信号模型实现源分离，主要包括声源距离模型、时频模型和序列模型等。Ozerov 等人[38]的研究则提出了基于结构化约束的多通道非负张量分解(Non-negative Tensor Factorization, NTF)[39]方法。该方法将多通道混合信号建模为三阶张量，通过引入谐波性约束和空间稀疏性约束，实现了用户可引导的音频源分离。这类基于模型的方法具有参数可调、物理意义明确等优势，能够较好地适应实际应用场景的需求。

近年来，基于深度学习的声源分离技术取得了突破性进展。早期的深度学习方法主要采用时频域处理策略，Hershey 等人[40]提出的深度聚类方法通过神经网络将时频点映射到高维嵌入空间，并结合聚类算法生成分离掩码，在单通道语音分离任务中取得了显著进展，标志着深度学习方法在该领域的有效性初步显现。为克服时频域方法存在的相位重构问题，研究者转向了时域端到端的解决方案。Luo 等人[41]提出的 Conv-tasnet 模型采用编码器、分离器、解码器架构，使用时域卷积网络(Temporal Convolutional Network, TCN)[42]直接处理波形信号，不仅在分离精度上超越了时频方法，还显著提高了运算效率。随着 Transformer 架构[43]在自然语言处理领域的成功，Subakan 等人[44]将 Transformer 架构引入声源分离任务，提出了 SepFormer 模型，推动声源分离技术实现了重要突破。该模型摒弃了传统依赖循环神经网络(Recurrent Neural Network, RNN)[45]的序列建模方式，采用时频双路径的多头注意力机制[46]，能够有效捕获语音信号中的短期与长期依赖关系。得益于 Transformer 的全局建模与并行处理优势，SepFormer 在多说话人语音分离任务中展现出优异性能，成为非 RNN 架构在该领域应用的代表性成果。

综上，声源分离技术经历了从传统的盲源分离与模型驱动方法到深度学习方法的演进，在分离精度和效率上取得显著进展，因此被广泛应用于语音信号处理任务。但该技术可能给音频水印带来新的安全挑战：一方面，攻击者可能利用此类工具对嵌入水印的音频进行攻击从而去除水印；另一方面，该项技术在各种音频信号处理过程中的使用易导致音频中的水印结构被“无意”改变。因此，设计具备抗声源分离攻击能力的音频水印算法具有重要的研究意义。

1.3 研究内容与结构安排

1.3.1 研究内容

本论文围绕抵抗声源分离攻击的鲁棒音频水印算法展开研究，主要基于深度学习工具构建音频水印嵌入与提取框架。随着深度学习在计算机视觉、音频信号处理及自然语言处理等领域的广泛应用，其在数字水印技术中的研究价值也日益凸显，逐渐成为热点方向。基于上述背景，本文结合不同神经网络结构的优势，针对声源分离攻击提出了两种神经网络音频水印算法。具体工作如下：

针对音频中声源分离攻击，提出了一种基于可逆神经网络的鲁棒音频水印算法。该算法利用可逆神经网络在信息隐藏技术中的优势，如前向网络与反向网络高度对称、水印嵌入和提取可由同一网络完成等特性，在音频的时域上嵌入水印。具体来说，为在音频嵌入水印过程中保留时间信息，首先对音频信号进行分帧。随后对水印进行调制并利用冗余编码技术将水印分散嵌入于音频信号中。通过对声源分离前后的失真进行分析，设计了模拟失真层。此外，在神经网络训练过程中，引入了低频损失函数，提高了水印的不可感知性。实验结果表明，该算法在对音频嵌入水印后仍拥有较高音频感知质量，且能抵抗声源分离攻击以及其它常见的攻击如噪声、滤波、重采样等。

考虑到时域嵌入水印对于声源分离攻击鲁棒性不足的问题，提出了一种基于生成对抗网络的频域音频水印算法。由于声源分离前后音频信号频谱特征中低频部分保留较完整，且人声信号更多存在于低频区域，该算法选择在音频特征中的低频区域进行水印嵌入。首先通过短时傅里叶变换从音频信号中提取频域特征，

接着对水印进行编码至与音频特征相同的维度，再对编码后的水印进行掩码操作，选择性地将水印嵌入音频特征中。在该过程中，引入水印嵌入强度因子，通过控制水印与信号的相对关系以实现水印在不可感知性和鲁棒性之间的平衡。实验表明，相较于时域水印算法，该算法在声源分离场景下拥有更好鲁棒性的同时，保证了音频在嵌入水印后的感知质量。且在面对其它常见信号攻击时，该算法同样具备良好的鲁棒性。

1.3.2 论文结构安排

本论文总共分为 5 个章节，其中各章内容结构如图 1.2 所示。

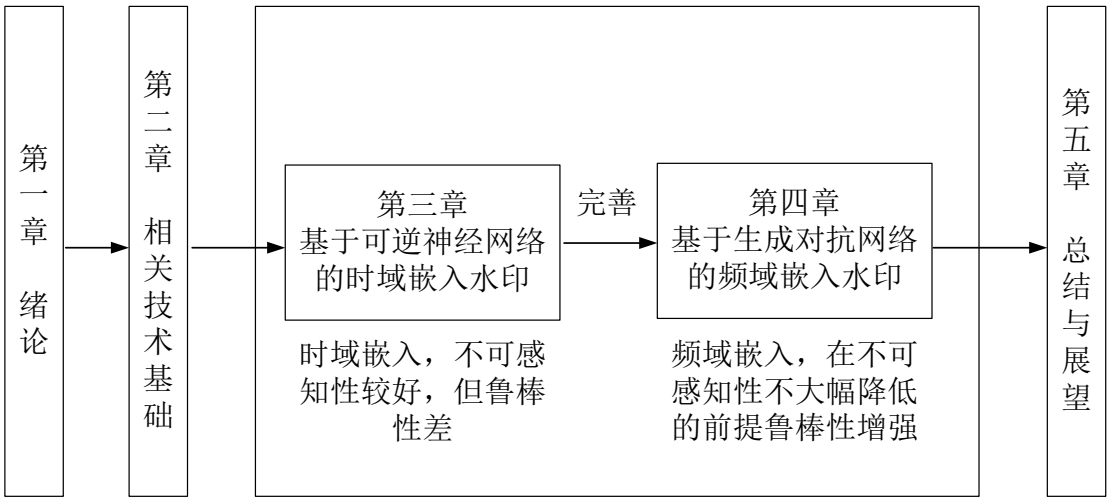


图 1.2 本文各章内容及其逻辑关系

第一章阐述了音频水印的研究背景和抗声源分离攻击的音频水印的研究意义，并着重介绍了音频水印技术以及声源分离技术的国内外发展现状。

第二章首先介绍了音频水印算法的基本框架以及音频水印算法的常用评价指标，随后介绍了声源分离技术以及其可能引发的安全威胁。其次，介绍了短时傅里叶变换技术。最后，对神经网络的基本概念以及各种网络结构进行了介绍。以上内容为后续研究提供了理论支持。

第三章为基于可逆神经网络的时域音频水印算法，首先介绍了该水印算法的整体流程以及可逆神经网络的训练过程。然后对声源分离过程中引入的失真进行了建模，设计了训练过程中的损失函数。最后，对该水印框架的算法性能进行了实验评估，包括水印的不可感知性，以及面对各种常见攻击和声源分离攻击的鲁棒性。

第四章为基于生成对抗网络的频域音频水印，阐述了所提出算法的整体框架，介绍了水印的嵌入与提取流程以及在训练过程中使用的损失函数。然后对实验中的参数设置进行了介绍。随后对水印的隐蔽性和嵌入容量进行了实验。最后，对水印面对各种常见电子攻击以及声源分离攻击的鲁棒性进行了评估，并针对嵌入强度因子和不同的失真层进行了消融实验。

第五章为总结与展望，对本文的研究内容进行了总结与概括，并对该领域的未来发展进行了展望。

1.4 本章小结

本章介绍了抗声源分离攻击的音频水印研究背景及研究意义，并介绍了国内外音频水印以及声源分离技术的研究现状，并在最后介绍了本文的主要研究内容与结构安排。

第二章　相关技术介绍

2.1　音频水印技术

2.1.1　音频水印框架

音频水印的基本框架如图 2.1 所示。其中核心模块包括水印嵌入模块和水印提取模块。水印嵌入模块负责将水印嵌入到载体音频中，使得水印在音频中无法被人类听觉系统察觉。水印检测模块则是从已嵌入水印的音频提取出水印信息。对于不同的应用场景，音频水印算法的鲁棒性有不同的要求。当面对版权保护、内容溯源等场景时，通常要求水印算法拥有良好的鲁棒性，在面对信道失真或其它失真后仍然能准确提取水印，此时称为鲁棒音频水印算法；在面对篡改检测等场景时，通常要求水印算法不具备鲁棒性，使得含水印的音频在经过修改后，事先嵌入的水印不能被正常提取，以此确定内容是否被篡改，此时称为脆弱音频水印算法。

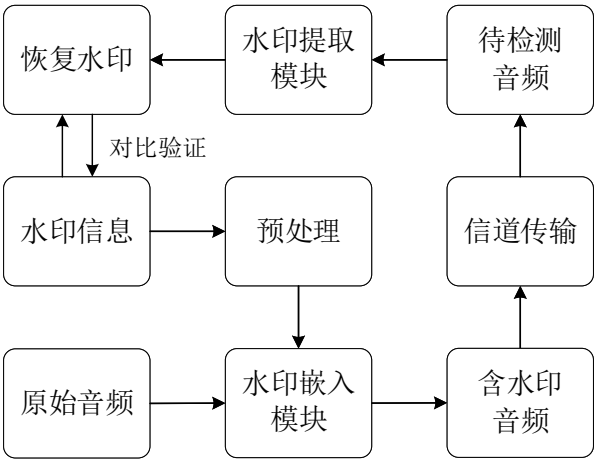


图 2.1　音频水印基本框架

预处理模块主要对载体音频和水印进行预处理操作，方便在后续过程中进行水印嵌入。对音频的常见预处理操作有帧划分、归一化等处理；对水印的预处理操作常见的有编码、加密等操作。嵌入的水印信息通常是包含版权信息的各种类型数据，可以是一段二进制序列、一幅图像或一段语音信号。载体音频和水印信息在经过预处理后，会一同被输入水印嵌入模块。水印嵌入模块采用特定算法，

利用载体音频的冗余特性，在不被发觉的前提下，将水印信息有效嵌入音频中。音频在嵌入水印后会在信道中进行传输。水印提取模块与水印嵌入模块相对应，通过水印提取算法对嵌入水印的音频进行处理，识别并恢复出原始水印。水印系统根据是否需要原始音频被分为盲水印和非盲水印两种，若需要原始音频参与水印提取，称为非盲水印算法；若不需原始音频参与水印提取，则称为盲水印算法。

随着深度学习的快速发展，许多基于深度学习的水印算法被提出，此类算法的核心框架由水印嵌入网络和水印提取网络组成。为了提高水印的鲁棒性，基于深度学习的音频水印算法在训练神经网络时会添加模拟失真层，对信号在真实世界遭受的各种失真进行模拟，引导水印提取网络学习如何从受干扰音频中提取水印。基于深度学习的音频水印的基本框架如图 2.2 所示。

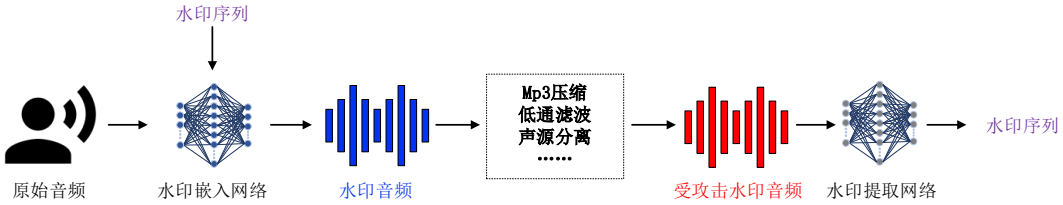


图 2.2 基于深度学习的音频水印框架

基于深度学习的音频水印算法可分为以下几类。第一种是基于编解码网络[34]以及水印提取网络的结构，其中编码网络负责将音频信号以及水印编码到特征空间中进行特征融合，利用解码网络将音频特征恢复成原始特征，最后通过相应的处理恢复原始信号。在提取水印时，将音频信号输入到训练好的神经网络，该网络会根据嵌入阶段所学习的特征映射关系，从输入信号中恢复出所嵌入的水印信息。第二种是基于生成对抗网络(Generative adversarial network, GAN)[35]的音频水印结构，基于 GAN 网络的音频水印在上述编解码网络结构上添加鉴别器，负责鉴别原始音频和添加水印的音频，通过引入对抗训练增加水印的不可感知性。第三种方法为基于可逆神经网络[33,47]的音频水印算法。与前两种采用编解码结构的水印算法不同，可逆神经网络在水印的嵌入与提取过程中使用完全相同的网络结构与参数。具体而言，可逆神经网络利用正向传播将水印信息与原始音频信号进行融合，生成含有水印的音频信号；而在提取阶段，通过网络的逆向结构，从水印音频中恢复出水印信息。可逆神经网络将音频与水印映射到共享特征空间中，

并保持映射的可逆性，因此在训练过程中，只需进行一次前向与反向传播，即可优化网络参数并实现嵌入与提取的联合建模。

2.1.2 音频水印评价指标

音频水印算法的整体性能通常可以从以下三个角度进行评价，有效载荷、鲁棒性和不可感知性，具体衡量方法如下。

有效载荷：有效载荷是指水印算法在不影响载体感知质量的前提下所能嵌入的最大信息量，是衡量音频水印算法性能的重要指标。在音频水印算法中，有效载荷通常以两种方式进行描述：

1）比特数

当每段音频仅嵌入单个完整的水印标识信息时，容量可表示为该段音频中所嵌入的水印总位数。例如，若在一段音频中嵌入了长度为 64 位的水印信息，则该音频的水印容量为 64 比特。

2）比特每秒

当水印以连续方式嵌入至整段音频中时，可使用比特率衡量水印的嵌入速率，表示系统每秒可嵌入的水印位数。该指标常用于实时水印系统的性能评估，其单位为比特每秒(bits per second, bps)。

本文对于所有的音频都嵌入相同长度的二进制比特水印，因此选择比特数作为本文水印有效载荷的评价指标。

不可感知性：不可感知性是在设计音频水印算法是首要考虑的指标。音频水印虽然利用音频信号中的冗余进行水印嵌入，但任何水印的添加都会造成原始信号的破坏，从而影响听觉质量，因此音频算法必须具有良好的不可感知性。常见的不可感知性评价标准可以分为主观评价和客观评价。

1）主观评价指标

主观评价是指从人的角度出发，根据主观评价标准表中的五个评分等级进行主观评价打分，取平均值作为最后的得分。这种方式非常直接，但是主观性较强。平均意见得分(Mean Opinion Score, MOS)值[48]是一种基于主观的评价方法，其评价标准如表 2.1 所示。MOS 值的评价方法为：让不同的测试者试听嵌入水印后的音频，然后根据音频的听觉质量对其进行打分，最后通过求测试者的平均分评定

音频听觉质量。MOS 值有 1~5 五个分值，5 分的质量最高，相当于音频几乎没有损失，4 分的音频质量稍差，但不影响听感和理解，3 分则存在明显失真，但音频的内容仍然可以理解，整体处于可接受水平，2 分则存在较严重的失真，理解上存在一定困难，且影响正常听感，而 1 分的音频音质极差，无法理解内容或无法忍受，严重影响听觉体验。

表 2.1 MOS 值对应音频评价标准

MOS 值	质量	描述
5	极好	音质极佳，几乎无法察觉失真
4	好	有轻微失真，但总体清晰可懂
3	一般	明显失真但不影响理解
2	差	严重失真，理解有一定困难
1	极差	无法理解或极其嘈杂，难以接受

2）客观评价指标

客观评价音频水印算法不可感知性的常用评价标准是信噪比(Singal to Noise Ratio, SNR)以及感知语音质量评估(Perceptual Evaluation of Speech Quality, PESQ)[49]。信噪比适用于各种音频信号，而客观语音质量评估更适用于语音信号。

信噪比是评估水印不可感知性的重要客观评价指标之一，其单位为分贝(dB)。信噪比通过计算信号能量与噪声能量的比值，反映水印嵌入过后音频失真的程度。嵌入水印过后音频的 SNR 值越高，则说明噪声能量越低，原始音频和嵌入水印之后的音频之间差别越小，则水印的不可感知性越高；反之，若 SNR 值较低，则说明水印的不可感知性较差。信噪比的计算公式如式(2.1)：

$$\mathrm{SNR}=10\log_{10}\left[\frac{\sum_{i=1}^{N}s^{2}(i)}{\sum_{i=1}^{N}(s(i)-s'(i))^{2}}\right] \tag{2.1}$$

其中 $s(i)$ 和 $s'(i)$ 分别表示原始音频信号和嵌入水印后的音频信号。

感知语音质量评估是另一种常用的客观评价指标，尤其适用于评估语音类音频在嵌入水印后的听感变化。PESQ 指标基于人类听觉模型，通过模拟人耳对语音质量的感知方式，从时间对齐、频域映射、掩蔽效应等多个维度综合评估原始语音与失真语音之间的主观感知差异。PESQ 的计算过程如图 2.3 所示。其评分

范围通常为-0.5 至 4.5，得分越高表示音质越接近原始语音，即水印的不可感知性越强；反之，若评分较低，则说明水印嵌入对语音感知质量造成了一定影响。

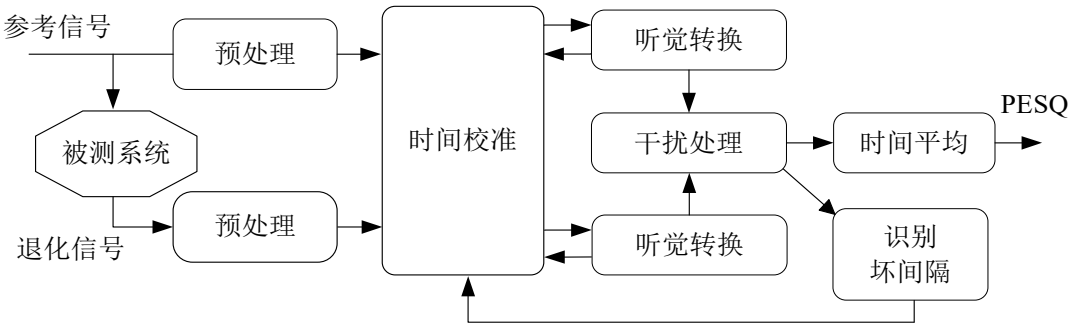


图 2.3 PESQ 计算流程

PESQ 作为 ITU-TP.862 标准被广泛用于语音通信、语音编解码和语音水印等领域的质量评价中，尤其适用于对语音清晰度和听感要求高的场景。PESQ 不仅弥补了 SNR 等纯数值指标无法反映主观听感的问题，也提供了更接近人耳感知的评估结果，适合用于语音水印系统中对不可感知性的评价。

鲁棒性：鲁棒性是指含水印音频信号在经过各种攻击，如压缩、裁剪、滤波、加噪等信号处理或其它攻击后，水印提取算法仍然能够从失真的音频中提取出水印进行认证。准确率(Accuracy, Acc)是衡量水印算法鲁棒性的一种常用评价指标，表示提取出的水印信息中与原始水印一致的比特所占比例。其定义为提取出的正确比特数与总比特数之比，用于反映算法在经过各种信号处理或攻击后的水印保留程度。准确率越高，说明水印在扰动条件下保持的完整性越好，鲁棒性越强。其定义如式(2.2)：

$$\mathrm{Acc} = \frac{B_{correct}}{B_{total}} \tag{2.2}$$

其中，$B_{correct}$ 表示正确提取的比特数量，B_{total} 表示总共嵌入的水印比特数量。Acc 越大，表明水印提取正确率越高，水印算法的鲁棒性性越好。

在设计音频水印系统时需综合考虑有效载荷、不可感知性与鲁棒性之间的平衡关系[50]。提高有效载荷往往会增加对音频的扰动，从而降低感知质量和鲁棒性；而更高的鲁棒性通常需要更高强度的嵌入方法，这会引起感知质量下降，也会压缩可用的嵌入容量；若需保证较高的感知质量，则必须控制嵌入强度，从而进一

步限制水印容量和鲁棒性。因此，在实际应用中，水印算法需根据具体任务需求，在三者之间进行权衡与协同优化，以实现最佳的综合性能。

除上述三个核心评价指标外，音频水印算法的评价指标还有诸如算法复杂度（时间复杂度和空间复杂度）、安全性以及实时性等。时间复杂度是指算法执行所需时间；空间复杂度是指算法运行时占用的存储空间，主要包含固定空间与可变空间两部分，固定空间是指算法本身占用的空间，可变空间是与输入规模相关的数据结构和递归调用等占用的存储空间；安全性是指攻击者知道水印存在的前提下，仍难以非法检测、提取或伪造水印的能力，其目标是防止攻击者利用其它算法恶意提取水印或破坏水印结构的完整性；实时性是指水印算法在短时间内完成嵌入与提取操作的能力，通常要求算法具有较低的延迟与计算开销，以满足在线处理、实时通信或流媒体传输等时效性需求。本文针对声源分离攻击提出鲁棒音频水印算法，故水印算法的鲁棒性表现成为首要关注点，而有效载荷和不可感知性作为水印系统的基础性能指标，也直接影响水印算法的实用性。因此，本文实验部分将集中分析这三个核心指标的表现，这既符合研究目标的需求，也能充分验证算法在目标场景下的适用性。

2.2 声源分离技术

声源分离是一项从混合音频中将不同声源信号分离出来的技术[51]，旨在从包含多个声源的音频中提取出特定目标声源。对于现实场景中存在多个声源的信号进行建模，单通道重叠声源信号可以表示为 n 个声源信号进行叠加，分别用 $x_1, x_2, x_3, \cdots, x_n$ 表示，叠加声源信号可表示为式(2.3)：

$$X = \sum_{i=1}^{n} x_i \tag{2.3}$$

其中 X 为重叠声源信号，x_i 表示第 i 个声源信号。声源分离问题可以表示为从混合声源 X 中估计每个声源信号 x_i，即得到 $\tilde{x}_i$，使得到的 $\tilde{x}_i$ 与 x_i 尽可能相似。

为了解决这一问题，早期研究提出了多种经典方法，主要基于统计信号建模。其中，独立成分分析假设各声源彼此统计独立，通过解混淆矩阵还原原始声源。非负矩阵分解则将音频的时频表示分解为若干基矩阵与权重矩阵，从而捕捉各声

源的频谱结构。随着计算能力的提升和大规模音频数据集的出现，基于深度学习的声源分离技术得到了迅速发展。2014 年，研究者首次提出基于深度神经网络(Deep Neural Network, DNN)的语音分离方法，相比于传统方法，该方法通过端到端的训练策略显著提升了性能。随后，考虑到语音信号的时间相关性，循环神经网络和长短时间记忆网络(Long Short-Term Memory, LSTM)[52]被广泛应用于处理序列数据，进一步改善了模型对语音动态变换的建模能力。近年来，卷积神经网络也被引入声源分离任务，用于提取语音的局部时频特征。其中，采用 U-Net 结构[53]的分离模型在提升分离精度和鲁棒性方面表现突出。随着 Transformer 架构的快速发展，研究者将自注意力机制应用到语音分离任务，使得模型在全局依赖建模能力方面进一步增强，取得了当前领先的性能。

尽管近年来声源分离技术取得了显著进展，但声源分离在处理音频信号的过程中，常常伴随着频谱重建误差、背景噪声残留等问题[54,55]。这些处理引入的非线性失真和结构性干扰会直接影响嵌入于音频中的水印信号的完整性与可提取性。例如，声源分离可能破坏水印嵌入位置处的时频结构，导致水印信息部分结构发生变化。此外，分离声源的过程中引入的噪声与滤波处理也可能削弱水印的能量，使其在提取阶段难以辨识。

2.3 短时傅里叶变换

在音频信号处理中，传统的傅里叶变换虽然能够直接将信号映射到频域，但它只提供了整体的频谱，无法直接反映音频信号在时间上的局部变化。而对于大多数音频信号（如语音、音乐），它们往往具有非平稳的特征，信号的频谱随时间不断变化。此时，如果仅使用傅里叶变换，就会把整个音频在频域上“摊平”，无法得到每个时刻对应的频率特征分布，也就失去了关键的时间信息[56]。因此，短时傅里叶变换引入了“分帧+加窗”的思路：首先将音频信号在时域上分成若干小段，每段称为一帧。对每一帧单独进行傅里叶变换，得到独立的频谱，再通过这些帧按时间顺序堆叠起来，就可以得到随时间变化连续的时频谱。

短时傅里叶变换主要分为以下两个步骤：

1）对音频信号进行分帧加窗处理。以窗口长度 N 和帧移 H 为参数，将音频序列分割为多个长度为 N 的短时帧，定义如下：

$$x_m[n] = x[n + mH], \quad 0 \le n < N \tag{2.4}$$

其中，n 为样本点，H 表示帧移，m 表示帧索引，N 表示每一帧的长度。为了减少帧边界不连续造成的频谱泄漏问题，对每一帧信号施加窗函数进行加窗。常用窗函数包括汉明窗(Hamming)、汉宁窗(Hanning)、布莱克曼窗(Blackman)等，优先使用汉明窗，其形式如式(2.5)：

$$w[n] = 0.54 - 0.46\cos(\frac{2n\pi}{N-1}) \tag{2.5}$$

经过加窗后，每帧变为：

$$x_m[n] = x[n + mH] \cdot w[n] \tag{2.6}$$

2）对每一帧信号进行快速傅里叶变换，将时域信号转变为频域信号。每帧加窗后得到长度为 N 的向量，执行 N 点快速傅里叶变换并在时间维度上进行拼接，其形式为：

$$X_m[k] = \sum_{n=0}^{N-1} x_m[n] \cdot e^{-j2kn\pi/N} \tag{2.7}$$

其中，$x_m[n]$ 为经过分帧加窗出以后的时域信号帧，N 为 FFT 的点数，$X_m[k]$ 为得到的复数谱，且满足 $0 \le k \le N-1$，包含幅度谱和相位谱。

2.4 深度神经网络

2.4.1 卷积神经网络

作为深度学习[57]领域的一种重要基础结构，卷积神经网络(Convolutional Neural Networks, CNN)[58-60]是一类专门用于处理具有网格状结构数据的前馈神经网络，在图像识别、音频信号处理以及自然语言处理等多个领域均取得了出色的表现。CNN 的核心思想主要包括局部连接与权值共享。局部连接是指每一层网络的神经元仅与上一层网络中的部分神经元相连，这一定程度上降低了网络的复杂度；权值共享则意味着同一个卷积核中的参数在整个特征图上保持一致，从而显著减少了模型的参数量。具体而言，CNN 通过使用具有局部感受野的卷积

核窗口，在特征图上不断滑动并进行卷积操作，从而有效地捕捉输入数据中的局部特征信息。此外，为了进一步增强网络的表达能力，通常需要设计多个卷积核，以便提取更加丰富且多样化的特征表示。

CNN 的结构一情况下都会包括输入层、卷积层、池化层和输出层，某些结构还会额外设置激活函数层、归一化层和全连接层，如图 2.4 所示。卷积层主要用于对输入数据进行特征提取，而池化层则通过对特征图进行下采样减少参数量，有效降低网络的过拟合风险，同时保留关键信息提高模型的鲁棒性。在实际网络中，卷积层与池化层通常交替出现。激活函数作为神经网络中的非线性变换单元，为网络提供强大的表达能力，使网络能够拟合更加复杂的数据分布。归一化层则通过规范化网络内部数据的分布，确保每一层的输入更加稳定，从而有效缓解训练过程中梯度消失或梯度爆炸的问题。

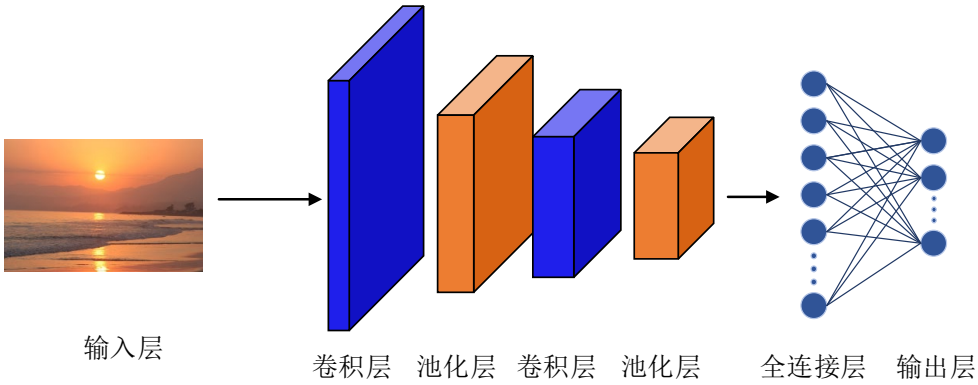


图 2.4 卷积神经网络结构图

卷积层是 CNN 的核心组件，其功能在于通过卷积核进行局部特征提取。卷积核本质上是一个可学习的权重矩阵，其关键参数包括尺寸、步长和填充方式。其中，卷积核尺寸决定了感受野的大小，直接影响特征提取的范围；步长控制卷积核的移动间隔，步长越大则输出特征图尺寸越小；填充操作（通常采用零填充）通过在输入边缘补零，既能维持特征图的空间维度，又能有效保留边缘信息的完整性。这些参数的合理配置对于网络的特征提取能力至关重要。以图 2.5 为例，其中输入特征图大小为 2×2，卷积核大小为 2×2，步长设定为 2，填充步长设置为 1，偏置为 1。卷积计算过程为从特征图左上角开始，按顺序选取与卷积核大小相同的矩阵区域，与卷积核进行逐元素相乘并进行求和，并叠加固定偏置，即

可得到该区域的特征值，将所有计算得到的特征值进行拼接，便得到最终的输出特征图。

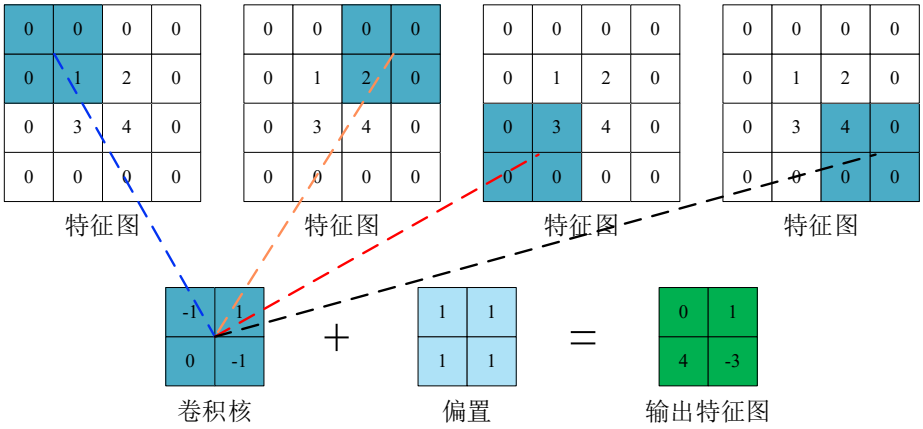


图 2.5 卷积计算过程图

池化层是 CNN 结构中的重要组成部分，位于卷积层之后，通常用于对卷积层输出的特征图进行下采样，以降低特征的维度，保留重要的特征信息，同时减少网络计算复杂度并提高模型鲁棒性。池化操作可以分为最大池化和平均池化两种方式。最大池化通过对池化窗口内的所有元素取最大值强化并突出局部区域内最明显的特征。平均池化在池化窗口内取所有元素的平均值，能够使得特征图更加平滑。图 2.6 展示了这两种池化操作的实现过程，设输入特征图大小为 4×4，池化窗口大小为 2×2，且无边界填充，池化核沿着特征图左上角以固定步长执行滑动窗口遍历。此时，池化核将分别在特征图的左上、右上、左下、右下四个区域内进行池化运算，每个区域根据不同的池化方式输出一个特征值，共形成 2×2 大小的输出特征图。这两种池化方式均可有效降低特征图维度，突出关键特征，提升神经网络的泛化能力。

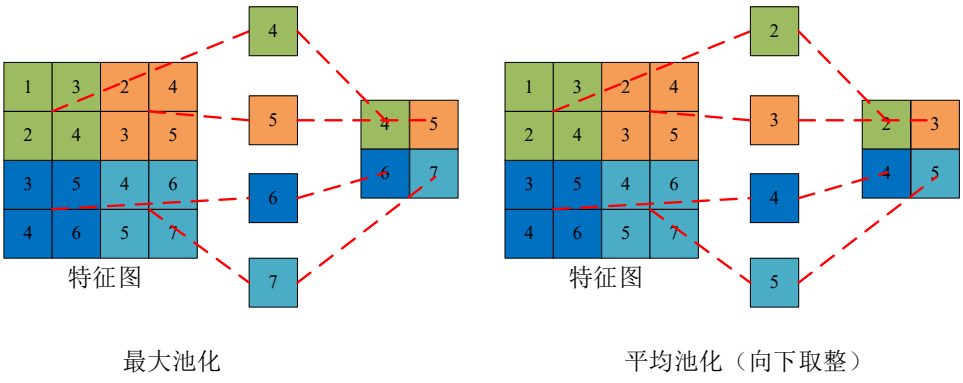


图 2.6 池化过程计算示意图

激活函数是卷积神经网络中一种非常重要的非线性映射工具，可以在网络中引入非线性因素，使得 CNN 能更有效的学习和拟合复杂的数据特征。根据函数特性差异，激活函数可以分为饱和型激活函数和非饱和型激活函数。饱和型激活函数是指当输入趋于无穷时，函数的输出逐渐接近某个稳定的极限值；非饱和型激活函数是指输入趋于无穷时，函数的输出不会逼近某个有限的极限值，而是无限增长。常见的饱和型激活函数有 Softmax、Sigmoid、Tanh 等，非饱和型激活函数有 ReLU、Leaky ReLU、Swish 等，其示意图如图 2.7。

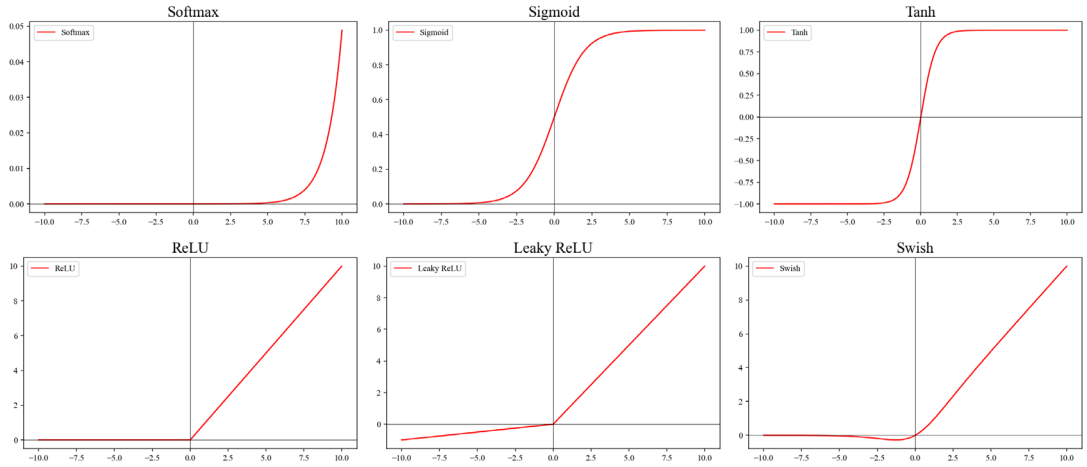


图 2.7 不同激活函数示意图

归一化层在深度神经网络中发挥着重要作用，不仅可以显著加快模型训练过程，还能有效缓解梯度消失和梯度爆炸问题，降低网络对参数初始化的敏感性，并增强模型的泛化性能。根据特征张量在批量大小 B 、通道数 C 、高度 H 和宽度 W 四个维度上的归一化方式差异，衍生出四种经典归一化方法：批量归一化(Batch Normalization，BN)、层归一化(Layer Normalization，LN)、实例归一化(Instance Normalization，IN)和组归一化(Group Normalization，GN)。BN 在每个通道上对由批量大小 B 、高度 H 和宽度 W 构成的维度中的特征计算均值与方差并进行标准化，因而适用于批量规模较大的训练场景；LN 在单个样本内沿 $C\times H\times W$ 维度整体标准化，通过消除特征维度间的统计差异，在循环神经网络和 Transformer 架构中表现出显著优势；IN 通过独立处理每个样本的单通道特征，仅基于 $H\times W$ 空间维度计算统计量，使其在生成对抗网络和图像风格迁移任务中广泛应用；GN 将通道划分为 G 个互斥子组，在每组内沿 $C/G\times H\times W$ 维度执行

标准化，其统计量计算与批量规模解耦，在小批量训练场景中具有良好的鲁棒性。各种归一化方法对特征数据的处理方式如图 2.8。

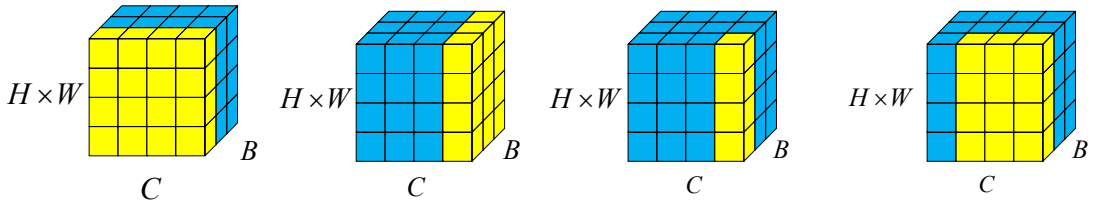


图 2.8 不同归一化方法示意图

残差网络(Residual Network, ResNet)[61]是深度学习中极具代表且影响深远的网络结构之一，其核心思想在于通过引入残差学习机制，使网络不再直接拟合目标映射函数 $H(x)$，而是学习输入与输出之间的残差函数，即 $H(x)=F(x)+x$。如图 2.9 所示，残差块通过引入跳跃连接(Skip Connection)，将输入信号直接加到经过权重层后的输出的 $F(x)$ 上，使网络更易学习到微小特征扰动，抑制了不良特征的影响。当目标映射趋近于恒等变换时，残差项可通过梯度下降快速收敛至零，显著降低深层网络的优化难度。在反向传播过程中，跳跃连接为梯度传播提供了一条无障碍的传递路径，有效缓解梯度消失问题，提升了模型的稳定性。

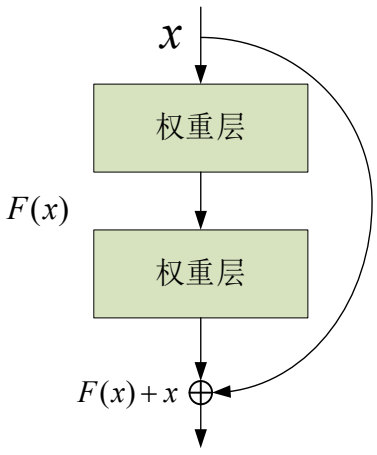


图 2.9 残差网络[61]结构图

稠密连接网络(Dense Connection Network, DenseNet)[62]由多个稠密块组成。在传统卷积神经网络中，每一层只与前一层相连，信息只能逐层传递。而在 DenseNet 的每个稠密块中，网络中的每一层都会接收前面所有层的特征图作为输入，从而使每一层在学习新特征的同时，还能访问所有前面层的特征，从而实

现了特征重用、更好的梯度传播和参数效率提升。密集连接的设计，极大提升了模型的表达能力、训练效率和参数利用率，稠密连接网络的结构如图 2.10 所示。

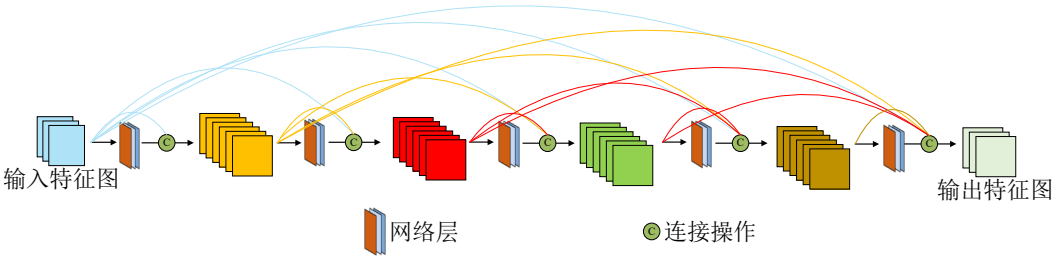


图 2.10 稠密连接网络[62]的结构

2.4.2 可逆神经网络

可逆神经网络[63,64]是一类特殊结构的神经网络，不同于传统前馈神经网络只关注输入到输出的单向映射，可逆神经网络可以实现输入与输出的双向映射，从而更好的保留信息。其基本思想是构建既可正向传播也可反向还原的映射函数。给定任意变量 x 和正向函数 $y=f_\theta(x)$，通过反向函数可得到 $x=f_\theta^{-1}(y)$。在此过程中，正向网络和反向网络共享相同的参数 θ。可逆神经网络的常见结构如图 2.11 所示，对于输入 x_1 和 x_2，可以由网络的前向过程得到 y_1 和 y_2。同样可以利用网络的逆结构，输入 y_1 和 y_2 得到 x_1 和 x_2。可逆块的公式如式(2.8)：

$$\begin{cases} y_2 = x_2 \odot \exp(\psi(x_1)) + \phi(x_1) \\ y_1 = x_1 \odot \exp(\eta(y_2)) + \rho(y_2) \\ x_1 = (y_1 - \rho(y_2)) \odot exp(-\eta(y_2)) \\ x_2 = (y_2 - \phi(x_1)) \odot exp(-\psi(x_1)) \end{cases} \tag{2.8}$$

其中，ψ、ϕ、ρ、η 为可学习的网络参数，$\odot$ 表示乘操作。

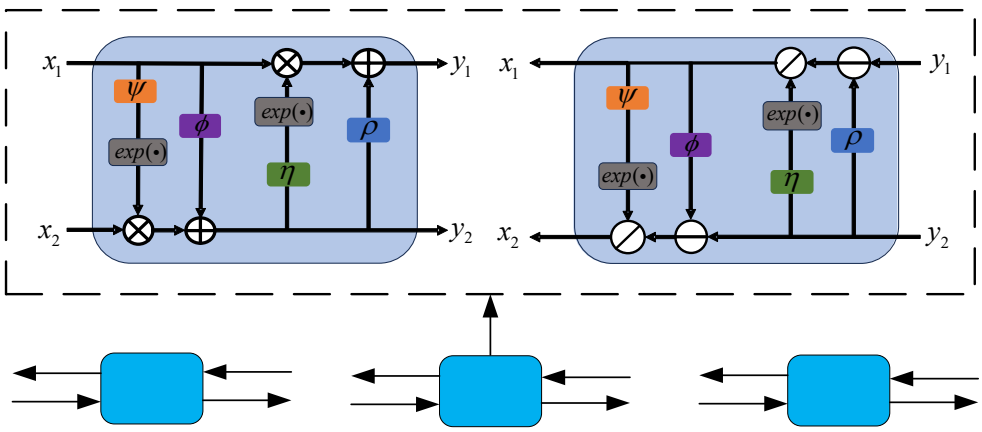


图 2.11 可逆神经网络[63]的结构

2.4.3 混合注意力模块

注意力机制作为一种模拟人类信息关注能力的仿生设计，可以在不显著增加计算开销的前提下，引导网络关注关键特征区域，从而提升模型在不同任务上的表现。卷积神经网络结构中，应用最为广泛的注意力机制主要包括：空间注意力、通道注意力以及空间与通道融合的混合注意力机制。

1）空间注意力：空间注意力机制旨在识别特征图对当前任务更为关键的空间区域。空间变换网络(Spatial Transformer Network, STN)[65]就是一种典型的实现方式。它在每个通道的 $H \times W$ 平面上根据特征中不同位置之间的相关性，为各位置分配不同的权重，从而引导网络关注更重要的空间位置。STN 通过仿射变换将特征图映射到新的空间，其结构主要包括三个部分：定位网络、网格生成器和采样器。局部网络通过卷积神经网络预测输入特征图的空间变换参数；网格生成器利用这些参数计算输入与输出特征图之间的坐标映射关系；采样器则根据该映射关系从原特征图中提取关键特征，生成变换后的输出特征图。其整体结构如图 2.12 所示。

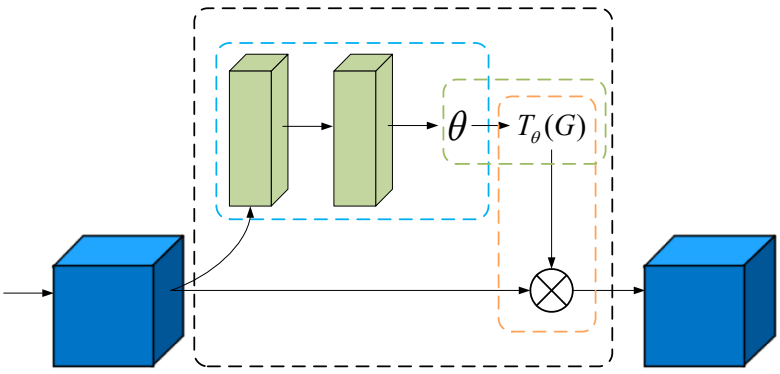


图 2.12 空间注意力模块

2）通道注意力：挤压激励网络(Squeeze-and-Excitation Networks, SENet)[66]引入了一种通道注意力机制，包括以下三个步骤。首先通过挤压操作，沿空间维度对每个二维 $H \times W$ 的特征图压缩，得到每个通道的全局描述向量。其次是激励操作，将这些通道的全局描述输入两层全连接网络，第一层根据压缩比进行降维，激活函数为 ReLU；第二层将特征向量升维回原始通道数。并使用 Sigmoid 函数对输出进行归一化，得到各通道的权重值，范围在 0 到 1 之间。最后再通过重定

标将这些权重乘回原始的每个通道的特征图，实现通道重要性的动态调节。SENet 的结构如图 2.13 所示。

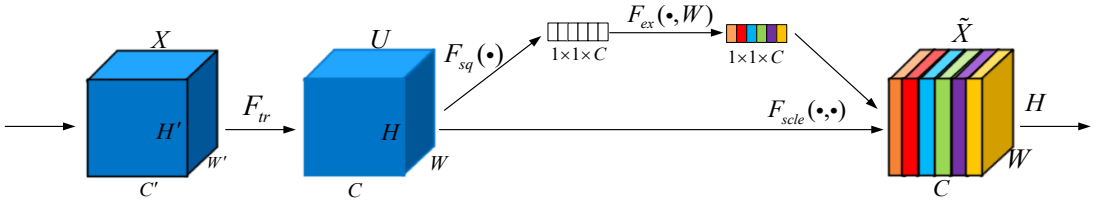


图 2.13 通道注意力模块

空间通道混合注意力：CBAM(Convolutional Block Attention Module)[67]是一种结构简洁、实现直观的空间与通道混合注意力机制。该模块先后引入通道注意力模块 M_c 与空间注意力模块 M_s，以此实现对通道特征和空间特征的联合建模与增强。在通道注意力模块 M_c 中，输入特征图 $F \in \mathbf{R}^{C\times H\times W}$ 分别经过全局池化与全局最大池化两个分支，得到两个大小为 $\mathbf{R}^C$ 的通道描述向量。随后，这两个向量输入共享权重的全连接网络，并进行相加与归一化处理，得到通道注意力权重。该权重与原始特征图逐通道相乘，得到通道增强后的特征图 F'。在空间注意力模块 M_s 中，对 F' 沿通道维度分别进行全局平均池化和全局最大池化，得到两个空间特征图后将它们合并后在经过卷积层和归一化操作后得到空间注意力图。最终，将空间注意力图与 F' 做逐元素乘法，得到增强后的输出特征 F''。CBAM 的具体结构如图 2.14 所示。

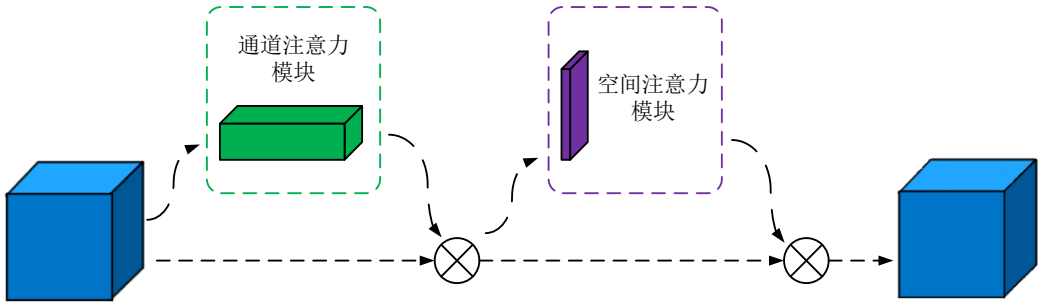


图 2.14 混合注意力模块[67]结构

2.4.4 生成对抗网络

生成对抗网络[68]的核心是通过两个神经网络之间相互博弈进行模型训练，它们分别是生成器和判别器。其中，生成器负责从潜在空间中采样噪声向量，通过非线性映射生成尽可能接近真实数据分布的伪造样本；而判别器则设计为二分类

网络，用以判断输入样本是真实样本还是由生成器伪造生成的。在训练过程中，判别器会输出 0 到 1 之间的概率值，代表当前样本是真实样本的概率。生成器的目的是生成尽可能真实的样本骗过判别器，而判别器的目的则是提升自身区分真实样本和伪造样本的能力。二者在对抗训练中不断迭代，最理想的状态就是判别器不能区分伪造样本和真实样本。由于其不依赖明确的概率密度函数建模，GAN 在图像生成、语音合成、风格迁移等任务中展现出强大的建模能力与生成质量，成为近年来生成建模领域的重要研究方向。GAN 网络的结构如图 2.15 所示。

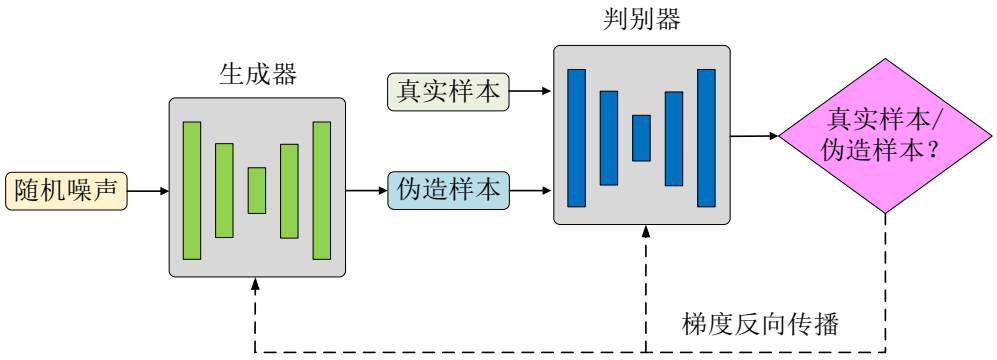


图 2.15 GAN 网络[68]结构图

2.5 本章小结

本章详细介绍了所提出方案的设计理论与研究基础。首先，系统阐述了音频水印相关技术，包括音频水印的基本概念、基本的水印嵌入与提取框架以及基于深度学习的音频水印框架。然后，对音频水印的评价指标进行了详细说明，涵盖了主观评价指标和客观评价指标，明确了评估水印不可感知性与鲁棒性的常用方法。接着，介绍了声源分离技术及其典型应用，分析了声源分离技术在实际应用中可能带来的安全威胁。最后，简要介绍了深度学习及神经网络相关知识，重点分析了神经网络中各层结构的功能与作用，为后续基于神经网络设计水印嵌入与提取模型奠定了理论基础。

第三章　基于可逆神经网络的时域音频水印

3.1　引言

音频水印是一种允许所有者在音频中隐藏其身份信息，便于后续进行身份认证的技术。鉴于音频知识产权保护的紧迫性和重要性，音频水印算法的研究从上世纪 90 年代就开始陆续提出，并且在研究的过程中不断的改进，出现了一系列经典的水印算法[69,70]。近年来，神经网络因为其强大的特征表示能力与非线性建模能力，逐渐进入信息安全相关研究者的视野。已有研究尝试采用卷积神经网络或自编码器框架在音频或语音信号中实现水印嵌入与提取，取得了初步成效。然而，这类方法多基于不可逆的编码结构，在信息嵌入与提取过程中容易丢失信号细节或引入冗余扰动，不利于提升整体系统的稳定性与信息保真度。

可逆神经网络的概念最早由 Dinh 提出[63]，该结构因其信息守恒特性和出色的重构能力，在各类信息隐藏与多媒体处理任务中受到广泛关注，并已被成功应用于图像隐写、可逆数据隐藏等领域[47]。应用到音频水印任务时，其可逆性结构能够在嵌入水印信息的同时最大程度地保留原始音频信号的细节特征，从而有效提升水印的不可感知性。此外，此类结构具备精确执行逆向映射的能力，能够在不借助原始音频的前提下从受扰音频中准确恢复嵌入的水印，提高了水印系统的鲁棒性。基于上述分析，本章提出一种基于可逆神经网络的时域音频水印嵌入与提取框架。相较于其它类型神经网络的水印方案，所提出方法在提升水印隐蔽性的同时，在水印提取的精度上也具备一定优势。

3.2　基于可逆神经网络的时域音频水印方案

图 3.1 展示了基于可逆神经网络的时域音频水印算法总体框架，主要由音频分帧模块、水印编码模块、可逆嵌入模块(Invertible Embedding Module, IEM)和重叠相加模块构成，整个系统在时域上实现水印与宿主音频的融合。使用音频分帧模块对音频进行分帧，将水印进行编码后与分帧后的音频一同利用可逆神经网络的前向过程嵌入水印，并通过重叠相加模块将音频恢复成一维信号。在提取水印

时，将嵌入水印的音频进行分帧操作并借助一个维度相同的随机噪声，一同输入可逆神经网络的逆向结构完成水印提取。

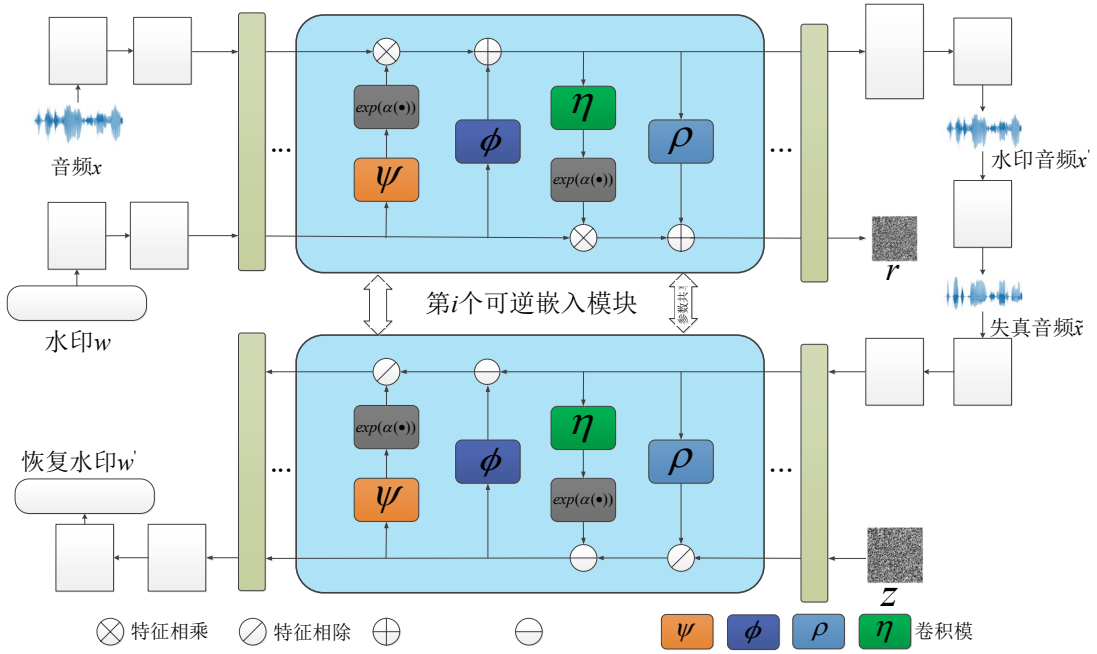


图 3.1 基于可逆神经网络的音频水印框架

3.2.1 水印嵌入与提取流程

本章算法中，将水印信号与音频的时域信号作为输入，使用可逆神经网络的前向传播嵌入水印。在前向传播过程中，首先将音频信号 x 进行分帧处理，得到音频矩阵 x_m，该操作如式(3.1)所示：

$$x_m = Enframe(x) \tag{3.1}$$

其中，*Enframe* 表示分帧函数，其作用是将一维音频信号按时间顺序排列成二维矩阵，确保水印嵌入过程中保留时间维度信息。接着，将待嵌入的水印比特序列输入至线性编码模块进行维度映射与冗余扩展，生成与音频矩阵维度一致的水印矩阵 w_m，如式(3.2)所示：

$$w_m = Repeat(Linear(w)) \tag{3.2}$$

其中，w 表示水印信号，*Repeat* 表示重复操作，*Linear* 表示线性神经网络，此处用于对水印进行扩频编码。随后，将音频矩阵 x_m 和水印矩阵 w_m 一同输入至可逆嵌入模块 IEM。IEM 由 N 个带参数的可逆嵌入块(Invertible Embedding Block，

IEB)组成，每个 IEB 中包含一个仿射耦合结构，每个仿射耦合结构由四个可学习函数实现特征提取与融合。函数能够自适应选择适合嵌入水印的音频区域。水印嵌入的过程中，共有 N 个 IEB 进行水印嵌入，第 i 个可逆嵌入模块中，输入输出可由式(3.3)至(3.4)表示：

$$x_m^i = x_m^{i-1} \odot (exp(\psi(w_m^{i-1}))) + \phi(w_m^{i-1}) \tag{3.3}$$

$$w_m^i = w_m^{i-1} \odot (exp(\eta(x_m^i))) + \rho(x_m^i) \tag{3.4}$$

其中 x_m^i 和 x_m^{i+1} 表示第 i 个可逆嵌入模块的输出和输出的音频矩阵，w_m^i 和 w_m^{i+1} 表示第 i 个可逆嵌入模块的输出和输出的水印矩阵，ρ、η、ϕ、ψ 表示可学习网络函数，$\odot$ 表示乘操作。可逆神经网络中的可学习网络函数并不要求可逆，因此 ρ、η、ϕ、ψ 无需遵守可逆变换的限制。为了更有效地从音频矩阵中提取深层特征并实现特征的重复利用，从而提高嵌入水印音频的感知质量和水印的鲁棒性，本章采用稠密连接网络作为可学习函数。

经过最后一个 IEB 后，音频矩阵和水印矩阵充分融合得到含水印音频矩阵 x'_m，将 x'_m 进行重叠相加和，得到最终嵌入水印的音频信号 x'，损失信息 r 被直接丢弃，如式(3.5)所示：

$$x' = \mathrm{O}verlapadd(x'_m) \tag{3.5}$$

其中 O*verlapadd* 表示重叠相加操作。得到含有水印的音频信号 x' 后，将其输入失真层模拟声源分离过程，从而得到失真的含水印音频信号 $\tilde{x}$，如式(3.6)所示：

$$\tilde{x} = Distortion(x') \tag{3.6}$$

其中 *Distortion* 表示失真处理操作。进行水印恢复时，对于得到的失真音频信号 $\tilde{x}$，对其进行分帧得到含噪音频矩阵，如式(3.7)所示：

$$\tilde{x}_m = Enframe(\tilde{x}) \tag{3.7}$$

然后利用可逆网络的逆结构对频谱信号进行水印提取。这个过程中需要使用到与输入时域音频矩阵信号维度相同的随机噪声辅助水印提取。提取的过程中第 i 个可逆嵌入模块，输入输出可以由式(3.8)至(3.9)表示：

$$\tilde{x}_m^i = (\tilde{x}_m^{i+1} - \phi(z^i)) \odot (-exp(\psi(z^i))) \tag{3.8}$$

$$z^i = (z^{i+1} - \rho(\tilde{x}_m^{i+1})) \odot (-exp(\eta(\tilde{x}_m^{i+1}))) \tag{3.9}$$

其中，z^i 表示经过第 i 个可逆嵌入模块后从辅助随机噪声中恢复出的水印信号。$\tilde{x}_m$ 表示失真的音频频谱信号。经过第 N 个可逆嵌入模块过后，将得到的恢复出的水印编码信号经过水印解码层。这个过程可以由式(3.10)表示为：

$$w'_E = (z^N - \rho(\tilde{x}_m^N)) \odot (-exp(\eta(\tilde{x}_m^N))) \tag{3.10}$$

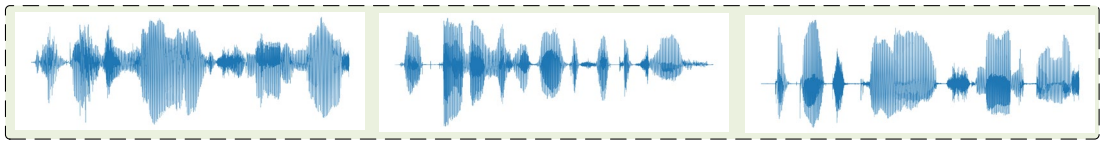
将得到的水印矩阵在与音频帧相同的维度上求平均值，然后通过与水印编码层相反的线性编码层进行水印提取，得到恢复的水印 w'，这个过程可以表示为式(3.11)：

$$w' = Linear(Average(w'_E)) \tag{3.11}$$

其中，*Linear* 表示线性神经网络，此处用于将水印信号调制回原始维度，*Average* 表示对水印矩阵取时间维度上的平均值。

3.2.2 声源分离失真建模

声源分离作为语音信号处理的重要技术，常被应用在背景去噪、伴奏分离、人工智能音频生成[71,72]等音频信号处理任务中。与传统攻击场景不同，声源分离攻击虽然在听觉感知层面引入的失真较小，但却可能破坏嵌入音频中水印结构，影响水印提取准确率。如图 3.2 所示，图 3.2(a)表示未分离的原始音频，图 3.2(b)和图 3.2(c)分别表示分离出的声源 1 和声源 2，从图中可以看出声源分离后的音频信号相较于原始音频信号在时域结构上引入了较为显著的差异，这种结构上的改动为音频水印带来了新的挑战。为解决上述问题，设计能够在声源分离处理后仍保持水印完整性的新型算法框架迫在眉睫。该类算法不仅需要具备良好的不可感知性并对常见电子攻击具备良好的鲁棒性，还需经声源分离后在得到的多个声源具备较高的水印恢复精度。



(a) (b) (c)

图 3.2 声源分离中引入的幅度失真：(a)未分离声源, (b)分离声源 1, (c)分离声源 2

为了增强所提出水印系统在声源分离攻击下的鲁棒性，本文在水印嵌入与提取之间进一步引入了失真模拟层(Distortion Simulation Layer)。该模块旨在训练过

程中模拟声源分离带来的失真现象，从而提升水印提取网络在此类攻击下的适应性。设计过程中参考了已有文献[73,74]对声源分离失真特性的研究，并结合实际应用需求，综合引入了以下三种失真方式：

1）高斯噪声(Random Noise, RN)：为了模拟声源分离过程中不可避免的背景噪声和残余噪声污染，本文在音频信号上叠加了信噪比为 20dB 的高斯噪声。通过调整信噪比范围，控制噪声强度，使得网络在存在轻微噪声污染的条件下仍能正确提取水印信息。

2）低通滤波(Low-pass Filtering, LF)：声源分离模型在分离过程中可能导致部分高频信息丢失。为此，本文采用低通滤波器对音频信号进行处理，截止频率被设置为 4000Hz。通过模拟分离器引起的频谱截断和高频衰减，从而提升模型对频率破坏类失真的鲁棒性。

3）幅度畸变(Amplitude Distortion, AD)：分离后音频常伴随局部幅度异常现象，如能量泄漏导致的局部放大或幅值压缩。本文通过对音频信号施加随机幅度缩放，具体操作对幅值进行放大或缩小 10%。通过模拟幅度畸变失真，使得水印系统能够适应在非均匀幅度变化环境下的提取任务。

通过上述失真模拟，训练过程能够更真实地反映声源分离过程中引入的失真效应，促使水印提取网络在复杂失真条件下学习到更加鲁棒的特征表达能力。

3.2.3 损失函数

音频水印模型的损失主要包括以下三种：保证嵌入阶段性能的嵌入损失、保证水印提取阶段性能的提取损失以及增强音频不可感知性的低频损失。三者在训练过程中共同对网络进行约束，以此改善模型的表现，提高水印的不可感知性和提取准确率。

嵌入损失 L_x 的目的是使得原始音频和水印充分融合，使生成的含水印音频在听觉上和原始音频难以区分。该损失定义如式(3.12)：

$$L_x = \mathrm{MSE}(x, x') \tag{3.12}$$

其中，x' 为嵌入水印后的音频信号，x 为原始音频信号，MSE 为均方误差函数。

提取损失 L_w 的目的是使得含水印音频在输入可逆神经网络逆结构后，得到

与原始嵌入水印一致的信息。该损失定义如式(3.13)：

$$L_w = \mathrm{MSE}(w', w) \tag{3.13}$$

其中，w'为恢复得到的水印信号，w为原始嵌入的水印信号。

由于在失真层中使用了低通滤波器，水印会较多地嵌入低频区域。因此，通过将原始音频和嵌入水印的音频进行离散小波变换，得到高频子带和低频子带，对音频的低频子带的水印嵌入强度进行约束，使嵌入水印后的低频子带与原始音频的低频子带尽可能相似，从而获得更好的不可感知性。该损失定义如式(3.14)：

$$L_{low} = \mathrm{MSE}(x_{ac}, x'_{ac}) \tag{3.14}$$

其中x_{ac}为原始音频信号的低频近似系数，x'_{ac}为水印音频信号的低频近似系数。

总体损失L_{total}是嵌入损失、提取损失、低频损失的加权和，其定义如式(3.15)：

$$L_{total} = \alpha L_x + \beta L_w + \gamma L_{low} \tag{3.15}$$

其中，α，β，γ是用来平衡这三种损失的超参数。

3.3 实验结果分析

3.3.1 数据集及实验设置

使用 Libri2Mix[75]作为实验的数据集，Libri2Mix 是 LibriMix 系列数据集中的一个重要组成部分，主要用于双人语音混合分离任务。该数据集通过将 LibriSpeech[76]数据集中的样本通过线性叠加操作将干净语音样本合成混合语音，以此模拟单麦克风下的混合声源。在训练时，将所有音频重采样到 16000Hz，每个音频的长度裁剪或补零至 2s。对于每个音频样本，分帧后的音频帧长度为 400 个采样点，相邻帧之间没有重叠部分。训练过程的迭代次数设置为 100，每次迭代过程打乱音频的顺序。损失函数的权重α，β，γ分别设置为 1，1，5。批大小(Batch Size)设置为 4，可逆模块的数量为 8，失真层的策略为每一个批次训练随机选中一种模拟失真或者不进行模拟失真，优化器为适应性矩估计(Adaptive Moment Estimation, Adam)，水印比特位数默认为 32，且$\beta_1 = 0.5$，$\beta_2 = 0.999$，学习率(Learning Rate, LR)固定为$10^{-4.5}$，随机种子统一设为 42。实验的硬件平台

为 Intel Core i7 10700 + RTX3090，CPU 主频为 2.9GHz，显存容量为 24GB。软件平台为 Python 3.8 + Pytorch2.4.1 + cuda12.2。

3.3.2 不可感知性分析

不可感知性是音频水印技术中的一项核心指标，指嵌入水印后对原始音频感知质量的影响程度。在理想情况下，嵌入水印后音频内容应当在听觉上对用户几乎无影响，从而其可用性以及用户体验不会受损。

表 3.1 含水印音频的不可感知性评估

评价指标	SNR(dB)	PESQ	MOS
文献[34]	26.12	3.50	4.38
本章算法	33.46	3.83	4.46

为全面评估本章所提出算法的不可感知性，采用主观评价指标(MOS 分)以及客观评价指标(SNR 和 PESQ)进行综合分析。实验在 Libri2Mix 数据集上进行，具体包括客观评价实验与主观听感测试两部分。实验计算了测试集中的音频在使用本章算法和文献[34]的算法嵌入 32 比特水印后的 SNR 值及 PESQ。用于对比的文献[34]为基于 DWT 的音频水印算法，实验所用参数设置与文献[34]保持一致。客观评价结果如表 3.1，文献[34]方法在 SNR 指标上达到 26.12dB，PESQ 指标结果为 3.50；而本章提出的算法在 SNR 达到 32.11dB，PESQ 得分为 4.21。可以看出，在本方法下，嵌入水印后的音频信噪比依然维持在较高水平，PESQ 指标也表现稳定，未观察到明显感知质量劣化，说明水印嵌入未对音频质量造成显著影响。主观评价方面，本文从 Libri2Mix 数据集中选取 50 个样本使用本章算法和文献[34]中的方法进行水印嵌入，并邀请 10 名受试者对原始及嵌入后音频进行 MOS 打分。文献[34]方法在 MOS 值为 4.38，本章方法则达到 4.46，整体得分优于文献[34]，表明在主观听感上，水印嵌入对音频质量影响极小，具有较高的不可感知性。综合主客观评价结果可见，本文提出的水印算法能够有效保证音频内容的感知质量。

为了更加直观地验证所提出音频水印算法在不可感知性方面的表现，在本节中进一步对水印嵌入前后音频信号的时域波形及其频域特征进行了对比分析。图 3.3 所示为原始音频与嵌入水印后音频的时域波形图，以及二者之间的波形残差

图。其中图 3.3(a)和图 3.3(b)分别为嵌入水印前后音频的时域波形图。由时域波形对比可观察到，嵌入水印后的音频信号在整体形态上与原始音频相比，未出现明显的结构性扭曲或局部突变现象。尽管在局部细节上存在微小差异，但该差异幅值极小且在时间轴上均匀分布，未造成任何形式的剧烈扰动，表明水印嵌入过程具有良好的隐蔽性。

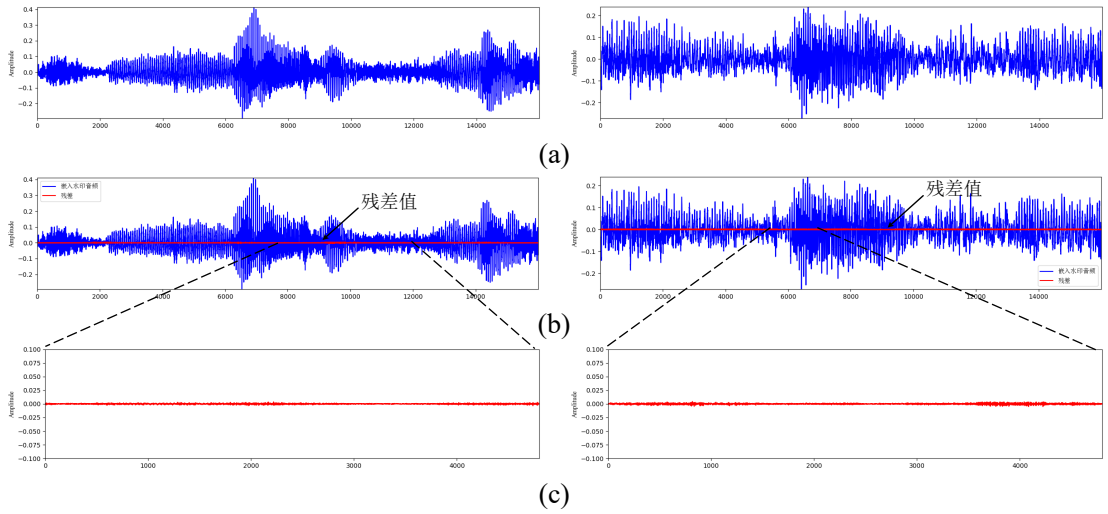
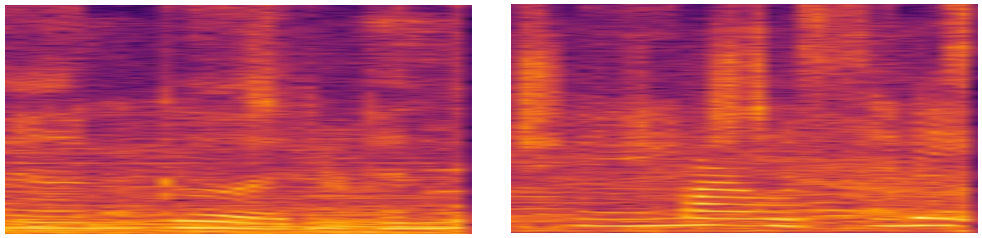


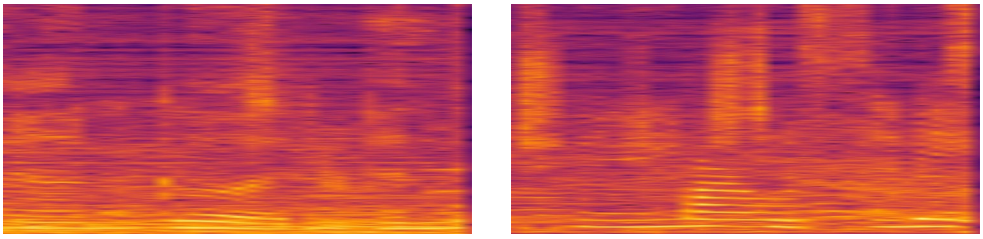
图 3.3 音频信号时域波形图：(a)水印嵌入前, (b)水印嵌入后，红色部分为残差值, (c)放大 10 倍后的局部残差值

此外，为进一步分析嵌入过程对音频信号造成的实际干扰，在图 3.3(c)中绘制了音频嵌入水印前后残差值放大十倍后的波形图，图中所示的时间长度为 0.3s。从图中可以看出，残差信号整体幅度较小，变化趋势平缓，与原始信号相比，波动幅度不超过音频最大幅度的 2%，进一步说明所嵌入的水印未对音频信号的局部结构或全局动态特性造成破坏。综合时域波形对比和残差信号分析结果可得，本文所提出的水印算法在保持音频感知质量方面具有显著优势，能够在不引入可感知失真的前提下实现水印信息的嵌入，从而满足数字音频场景中对高不可感知性水印的实际需求。

图 3.4 为嵌入水印前后音频频谱特征。其中图 3.4(a)和图 3.4(b)分别为嵌入水印前后的音频信号频谱图。从图中可以看出在频域方面，原始音频和水印音频的频谱分布几乎完全重合，频率成分的一致性得以良好保持。无论是低频还是高频段，嵌入水印后音频的频谱能量变化均非常有限，未出现频谱能量集中降低或频带失真的现象。这说明水印的嵌入不会对音频的频谱特征造成显著干扰，从频域角度进一步佐证本章水印算法良好的不可感知性。



(a)



(b)

图 3.4 音频嵌入水印前后频域上的对比: (a)嵌入水印前，(b)嵌入水印后

3.3.3 有效载荷分析

有效载荷作为音频水印的重要指标，用于描述单位音频中能够嵌入的水印信息量。在实际应用中，有效载荷的大小直接影响系统在传递、认证或标识信息方面的实用性。然而，提高有效载荷会导致音频质量下降或水印鲁棒性降低，因此在水印算法设计中，需要在有效载荷、不可感知性和鲁棒性三者之间进行权衡。

为全面评估所提出的音频水印算法在不同嵌入容量条件下的性能表现，本文设计了一组容量控制实验，旨在探索算法在提高信息承载能力的同时对音频质量与水印提取准确性的影响。选用 Libri2Mix 测试集进行实验，水印嵌入比特位分别设置为 16 位、32 位、64 位和 128 位。评价指标采用嵌入水印后音频的信噪比以及 PESQ 和水印提取准确率 Acc。实验的过程中不对音频进行其它信号处理，以更清晰地反映嵌入容量这一单一因素对系统性能的影响。

表 3.2 嵌入不同位比特水印时各项性能指标

有效载荷(比特)	SNR(dB)	PESQ	Acc(%)
16	35.38	4.12	100
32	33.46	3.83	100
64	30.15	3.72	100
128	28.89	3.51	100

如表 3.2 所示，随着水印容量的逐步增加，音频信号的 SNR 和 PESQ 指标有所下降，反映出嵌入过程对音频质量产生了渐进式的影响。在嵌入 16 比特水印时，水印音频拥有最高的不可感知性，其中 SNR 能达到超过 35dB，PESQ 评分也能保持在 4.0 以上，表明嵌入 16 比特水印不会对音频的感知质量造成较大影响。即使在嵌入水印达到 128 比特的情况下，含水印音频相较于原始音频仍能达到超过 28dB 的信噪比，且 PESQ 保持在 3.5 以上，虽然感知质量有所下降，但仍在高质量语音范围。同时，水印提取准确率始终维持在 100%，验证了所提出方法在嵌入较多水印时仍能完整提取。因此，该方法实现了对嵌入容量和不可感知性的有效平衡，具有较强的实用性。

3.3.4 水印鲁棒性分析

音频水印的鲁棒性是指嵌入水印的音频在面对各种干扰、攻击或噪声时，水印仍然能够被顺利提取。音频水印技术在实际使用过程中，含水印的音频信号可能在传输时经过信号处理或信道失真，而水印算法要保证在面对这些常见的处理或失真下依旧保持可靠性。因此，评估水印的鲁棒性表现时验证水印算法的关键一环。本节主要就音频水印的压缩编码、重采样、滤波以及噪声攻击进行讨论，使用水印提取准确率作为评价指标。最终取得的结果如表 3.3 所示。

表 3.3 不同算法在面对常见信号攻击时的表现(%)

攻击类型	文献[34]	本章算法
无攻击	100	100
Mp3 压缩(64kbps)	100	100
Mp3 压缩(128bps)	100	100
随机噪声(20dB)	96.22	98.12
8 比特量化	98.98	100
回声处理	100	97.19
低通滤波	99.97	99.06
重采样 1	96.22	97.03
重采样 2	97.33	95.12
中值滤波	99.98	99.30
幅度畸变	100	100

由表 3.3 可以看出，在无攻击条件下，本章算法与文献[34]均能够实现 100%的水印提取准确率，表明在理想环境下两者均具备稳定且无失真的水印嵌入与提取能力。针对 Mp3 压缩攻击实验，本章在 64kbps 和 128kbps 两种压缩强度下进行了测试，结果显示，无论是文献[34]方法还是本章算法，在两种压缩条件下的水印提取准确率均保持在 100%，说明在常见比特率下的有损压缩处理不会对两种方法嵌入的水印造成明显影响，水印信息能够可靠地嵌入和恢复。在噪声攻击实验中，本章算法在加入 20dB 白噪声条件下的提取准确率达到和 98.12%，而文献[34]方法较本文有 1.9%的准确率下降，表明本章所提出的算法在因对噪声类攻击时具有更好的鲁棒性。在回声处理攻击下，文献[34]的水印提取准确率维持在 100%，而本章所提出的算法准确率为 97.19%，尽管存在一定程度的下降，但整体仍处于较高水平，表明本章方法在面对典型的环境回声失真时，水印信息大部分能够成功提取，具备较好的稳定性。在低通滤波攻击条件下，本章算法的提取准确率为 99.06%，与文献[34]方法（99.97%）接近，表明即使在高频成分被削弱的情况下，本章算法依然能够有效恢复水印信息。此外，在重采样攻击测试中，无论是 8kHz 降采样重建（重采样 1）还是 32kHz 升采样重建（重采样 2），本章方法均取得了 97.03%和 95.12%的准确率，与文献[34]（96.22%、97.33%）取得了相当的性能表现，验证了本章算法在采样率变化环境下的鲁棒性。在中值滤波攻击下，本章算法达到了 99.30%的水印提取准确率，尽管略低于文献[34]的 99.98%，但整体表现仍然优异，显示出在局部时域扰动环境下良好的提取稳定性。最后，在幅度畸变攻击实验中，畸变的幅度为放大或缩小 10%，文献[34]与本章算法均实现了 100%的提取准确率，表明轻微的幅度变化对两种方法均无显著影响，验证了算法在时域幅度扰动下的稳健性。综合上述实验结果，本章提出的音频水印算法在多种典型攻击环境下均展现出良好的鲁棒性，特别是在噪声干扰、采样率变化等复杂失真条件下，相比文献[34]方法表现出更优异的稳定性与提取准确率。

3.3.5 水印在声源分离场景下的性能评估

为模拟常见的声源分离操作，本节实验采用 SpeechBrain[77]发布的预训练模型 SepFormer 以及 Conv-tasnet 模型作为声源分离工具，二者均在多个公开语音

分离任务中表现出色，具有较强的分离能力与泛化性。模型输入采样率为 16kHz，输入为单通道混合声源信号，输出为两路分离声源信号。实验从 Libri2Mix 测试集中随机选取 10 个样本作为基础数据源，采样率为 16kHz，每段音频长度约为 5 秒，内容覆盖不同性别与不同说话风格的语音，具备良好的多样性，其时域波形如图 3.5 所示。

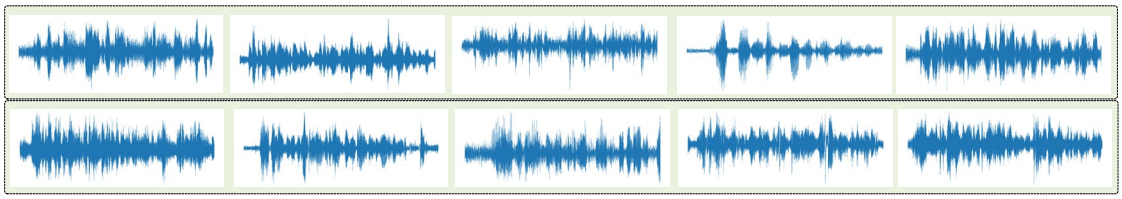


图 3.5 实验样本时域波形

实验首先将随机生成的 32 位二进制比特信息嵌入到原始音频中，生成含水印音频，再使用声源分离模型对含有水印的音频信号进行声源分离得到两个不同的声源信号。最后对这两个声源以及水印音频利用训练完毕的网络进行水印提取，采用水印提取准确率作为本节实验的评价标准。表 3.4 列出了在不同测试样本上水印音频以及分离后的声源水印提取准确率。

表 3.4 本章算法在不同声源分离模型攻击下的水印提取准确率(%)

样本	水印音频	SepFormer		Conv-tasnet	
		声源 1	声源 2	声源 1	声源 2
样本 1	100	100	62.50	81.25	59.38
样本 2	100	84.38	71.88	56.25	100
样本 3	100	96.88	62.50	56.25	59.38
样本 4	100	84.38	78.12	56.25	100
样本 5	100	68.75	90.62	65.62	90.62
样本 6	100	81.25	90.62	62.50	84.38
样本 7	100	93.75	59.38	96.80	56.25
样本 8	100	81.25	43.75	62.50	71.88
样本 9	100	93.75	81.25	65.62	65.62
样本 10	100	100	59.38	87.50	71.88
平均	100	88.44	70.01	69.05	75.94

由表 3.4 可以看出，本章水印算法在未经过声源分离处理的水印音频上，水印提取准确率始终保持在 100%，表明水印嵌入过程对音频内容的影响极小，同时水印信息能够完全恢复，验证了算法在无攻击环境下的良好性能。对于使用 SepFormer 分离后的声源 1，水印提取准确率略有下降，平均准确率为 88.44%。大部分样本能够维持较高的提取准确率（如样本 1、样本 10 达到 100%），但部分样本出现准确率下降（如样本 5 提取准确率不足 70%），表明声源分离过程中对水印结构造成了一定干扰。在声源 2 上，水印提取准确率有一定下降，平均准确率为 70.01%，并且整体准确率波动较大。其中个别样本（如样本 5、样本 6）仍能维持较高水平（90%以上），但多数样本的准确率低于 80%，反映出声源分离器在处理声源 2 时引入了更严重的失真，导致水印恢复效果受限。综合上述结果可见，声源分离攻击对音频水印的提取性能有显著影响，尤其是在次要声源（声源 2）中表现更为明显。推测主要原因在于：声源分离过程中存在能量泄漏与残余噪声，破坏了水印承载区域；分离出的两个声源在音频质量上存在差异，导致提取准确率不同程度下降。

为了验证算法的通用性，本章算法选取了另一种经典的声源分离网络 Conv-tasnet，并对嵌入水印的音频进行了相同的声源分离操作。在使用 Conv-tasnet 得到的结果中，声源 1 的平均水印提取准确率为 69.05%，而声源 2 的水印提取准确率为 75.94%。虽然对于声源 1 的水印提取准确率稍低，但平均水印提取准确率仍能达到 70%以上，因此在一定程度上对使用 Conv-tasnet 分离声源也具有一定鲁棒性。其中声源 1 中有个别样本提取准确率能到相对较高的水平(样本 1、样本 7、样本 10)，而声源 2 中有样本甚至达到 100%。但仍有部分样本提取准确率较低，如样本 3 和样本 9，其两个声源的水印提取准确率分别不足 60%和 70%。

综合对比可见，该音频水印算法在应对不同的声源分离模型时均具有一定的鲁棒性。在面对 SepFormer 分离声源时在声源 1 上表现了稍高的水印提取准确率，而声源 2 则相对较差。但在使用 Conv-tasnet 分离时得到的结果则恰恰相反。这种差异可能源于两种声源分离模型在建模方式、特征提取以及分离策略上的不同。总体而言，对于不同的声源分离模型，虽然水印提取准确率在分离出的声源信号上有所差异，但平均准确率都高于 70%，这表明该音频水印算法能够在一定程度上抵抗声源分离攻击。

3.4 本章小结

本章提出了一种基于可逆神经网络的鲁棒音频水印算法，该算法能在一定程度上抵抗声源分离攻击。为了在嵌入过程中保留音频的时间信息，首先对音频进行分帧得到音频矩阵信号，接着通过编码将水印编码至于音频矩阵相同维度，再将编码后的水印于音频矩阵共同输入可逆神经网络进行水印嵌入，然后通过重叠相加模块对含水印的音频矩阵进行恢复得到一维音频信号。在得到含水印的音频信号后，通过失真层模拟声源分离中的失真，并对失真的含水印音频信号再一次分帧后并借助辅助噪声输入至可逆神经网络逆结构进行水印提取。为了提高水印的不可感知性，本章引入一种低频损失，以实现水印音频的低频区域与原始音频的低频区域尽可能相似。实验结果表明，所提出的算法具有良好的感知质量和鲁棒性。算法对常见的如噪声、滤波、重采样等电子攻击具有良好的鲁棒性，音频在嵌入水印后能达到较高的信噪比以及较好的主观听感。此外，本章提出的基于可逆神经网络的音频水印算法对声源分离场景下也具有一定的鲁棒性。

第四章　基于 GAN 网络的频域音频水印

4.1　引言

随着音频水印技术的快速发展，如何在保证音频质量的前提下实现有效且鲁棒的水印嵌入，成为了音频安全领域中的重要问题。在上一章，提出了一种基于可逆神经网络的时域嵌入水印方法，该方法通过在时域上直接进行水印嵌入，有效地保证了音频的不可感知性，同时通过模拟失真增强了水印的鲁棒性，使得水印能够在常见的噪声干扰和攻击（如压缩、重采样等）下成功提取。然而，尽管该方法在大多数常规应用场景下表现良好，但在面对声源分离场景时，暴露出水印算法鲁棒性不足的问题。其原因在于声源分离过程会引入较大的幅度失真，水印信息可能被分离成不同的音频成分，导致水印提取准确度大幅下降，这成为时域水印算法的一大瓶颈。

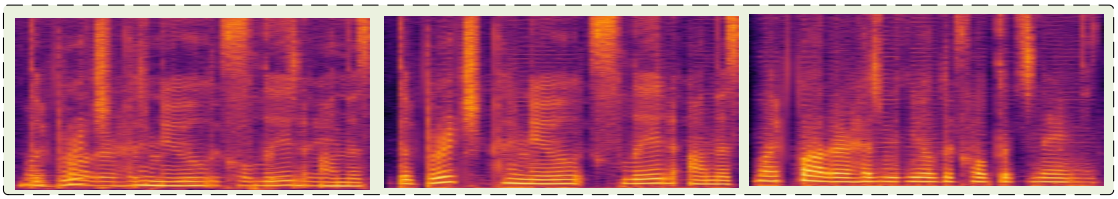


图 4.1　音频频谱图：(a)未分离音频信号, (b)分离声源 1, (c)分离声源 2

为了克服时域水印算法的局限性，提升水印在声源分离攻击场景下的鲁棒性，本章提出了一种基于频域的水印嵌入方法。相较于时域水印，频域处理能够更好地保留音频信号的全局特征，在一定程度上避免时域波形中的微小扰动对音频质量的影响。图 4.1 展示了声源分离前后音频信号的频谱对比。从中可以看出，经过声源分离后，声源 1 和声源 2 的频域特征仍得以较好的保留，尤其是在低频区域，大部分关键信息未被破坏。因此，本章选择将水印嵌入音频低频区域的频谱特征中，增强算法对抗声源分离攻击的鲁棒性。此外，本章进一步引入了生成对抗网络的框架，利用对抗训练的方式提升水印不可感知性[78]。通过对抗网络的优化，网络能够学习到更加隐蔽和鲁棒的水印嵌入方式，即使在声源分离攻击下，水印的提取准确性依然能够得到较好的保持。因此，本章重点介绍频域嵌入水印

方法的设计思路与网络结构，并在后续实验中验证了多个场景下水印的鲁棒性以及不可感知性。

4.2 基于 GAN 网络的频域音频水印方案

4.2.1 总体框架

本章提出了一种基于 GAN 网络的频域鲁棒音频水印，旨在在保证音频质量的同时提升水印在各种信号失真下的鲁棒性。该方法的总体架构如图 4.2 所示，主要包括三个不同的神经网络，分别为水印嵌入网络（生成器，用 G 表示），鉴别器（用 D 表示）、水印提取网络（解码器，用 E 表示）。为了提高鲁棒性，引入模拟失真层对真实世界中的信号失真进行模拟。每个网络不同的损失函数约束，从而实现不同的功能。其中，水印嵌入网络用于将水印信息嵌入音频信号的频谱特征中，进而由相应反变换转换为含水印音频信号；鉴别器用于区分含水印的音频信号和不含水印的音频信号；水印提取网络用于从嵌入水印的音频信号中提取水印进行验证。

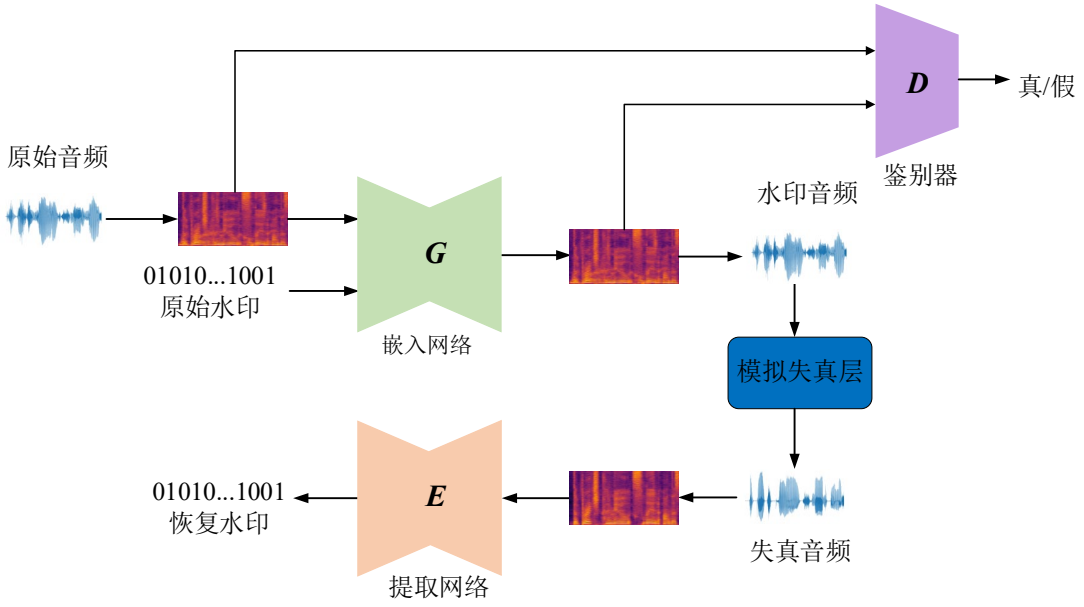


图 4.2 基于生成对抗网络的音频水印流程图

首先，对原始音频信号进行预处理与特征提取，同时将待嵌入的水印信号编码至与音频特征相同的维度，再对编码后的水印信号进行掩码，以减少对音频信

号频谱的修改。接着，将音频特征与编码后的水印信号共同输入水印嵌入网络，并通过不同的嵌入强度因子控制水印的嵌入强度，以生成包含水印信息的音频特征，再利用频域反变换重构为含水印音频。随后，计算含水印音频与原始音频之间的嵌入损失，用于保证水印嵌入前后音频的感知质量。进一步地，将含水印音频特征与原始音频特征输入鉴别器，以获得判别结果并计算鉴别器损失，从而引导嵌入网络提升水印的隐蔽性。为模拟实际场景中的信号畸变，含水印音频经过模拟失真层处理，再输入水印提取网络以恢复出水印信息，并据此计算水印提取损失，用于提高水印的鲁棒性。最终，通过联合水印嵌入损失、水印提取损失与鉴别器损失，对水印嵌入网络和提取网络进行协同优化，从而获得更具隐蔽性和鲁棒性的水印嵌入与提取模型；同时，利用鉴别器损失单独优化鉴别器网络，以提升其判别能力。

4.2.2 水印嵌入与提取流程

由于音频具有时间维度信息，在使用常见傅里叶变换时会将整个频谱均摊，导致时间信息消失。而短时傅里叶变换通过引入分帧和加窗的思想，能够较好地保留时间信息，因此本章使用短时傅里叶变换将音频变换到频域。在使用 2.3 节提到的短时傅里叶变换后，可得到音频的频域特征以及相位特征，该过程如式(4.1)所示：

$$c, p = STFT(x) \tag{4.1}$$

其中，x 为原始音频信号，$STFT$ 为短时傅里叶变换，c 为音频信号的频谱特征，p 为音频信号的相位特征。然后对水印进行如式(4.2)的扩频和重复编码，其具体操作为：对于一段输入的水印序列 w，采用线性神经网络和重复操作然后进行掩码操作，掩码策略为每十个时间帧范围内保留单个水印帧，其它区域设置为 0。得到编码后的水印 w_E，其维度大小与音频信号频谱图特征相同。

$$w_E = Mask(Repeat(Linear(w))) \tag{4.2}$$

其中，$Mask$ 为掩码操作，$Repeat$ 为重复操作，$Linear$ 为线性神经网络，此处用于对水印进行扩频编码。水印嵌入网络的目的是往原始音频信号的频域特征中嵌入水印信号 w，且需要满足在嵌入了水印后的音频信号 x' 在听觉感知质量上与

原始音频信号 x 接近，否则易引起攻击者的怀疑。水印嵌入网络结构如图 4.3 所示。

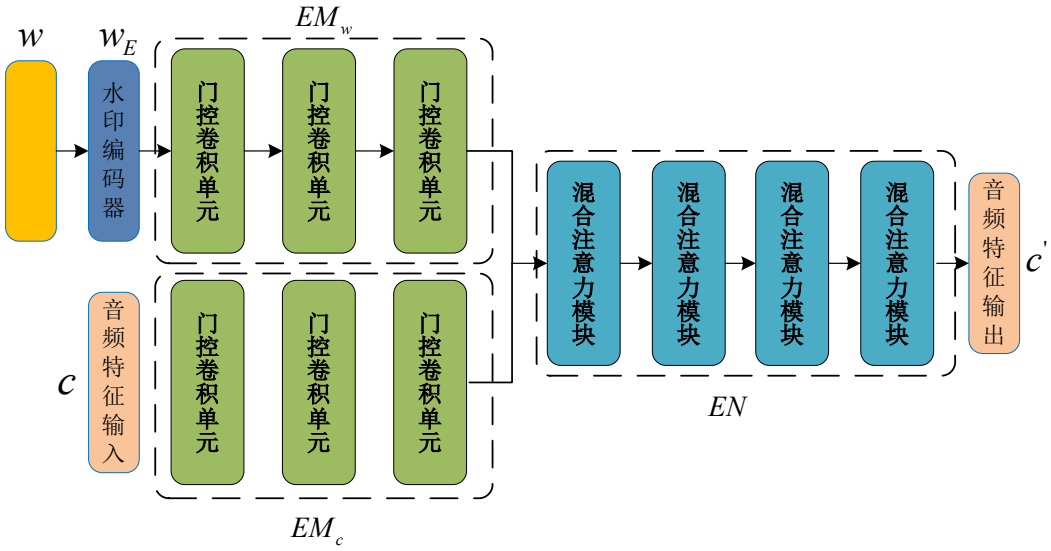


图 4.3 水印嵌入网络结构图

为了进一步提升网络对关键语音特征的感知能力，增强编码器中的特征表达效果，水印编码模块 EM_w 和音频特征编码模块 EM_c 均由三个门控机制的卷积单元(Gated Convolutional Unit，GCU)[79]组成，该模块借鉴了门控线性单元(Gated Linear Unit, GLU)在自然语言处理[80]与语音建模任务[81]中的有效性，通过引入门控结构对特征通道进行动态调节，以实现信息的选择性传递与抑制。GCU 的核心思想在于：在普通卷积操作后，将输出通道划分为两部分，一部分用于建模主信息，另一部分则通过 Sigmoid 函数生成门控系数，对主信息进行逐通道调节。与传统的 ReLU 或 PReLU 激活不同，GCU 通过学习门控函数动态决定每个通道是否“打开”或“关闭”，从而有效抑制冗余特征的传播，提升模型对关键信息的建模能力。水印信号和音频特征在经过编码过后拼接在一起输入到解码模块中。解码模块由四个 CBAM 模块组成，每个 CBAM 模块包含两个子模块，分别为通道注意力模块与空间注意力模块。通道注意力模块用于显式建模特征图中各通道的重要性，通过引入全局平均池化与全局最大池化两种统计方式分别从输入特征图中提取通道层级的描述信息，并共享一组使用 1×1 卷积实现的全连接卷积网络，将两种统计特征相加融合，利用 Sigmoid 函数生成通道维度的注意力权重

图，最后与输入特征通道相乘，实现对通道的增强。为避免在通道数较小时出现维度压缩至零的情况，通道数压缩比设置为式(4.3)：

$$c\_out = max(1, c\_in // r) \tag{4.3}$$

其中 c_out 表示输出通道数，c_in 表示原始通道数，$//$ 表示整除运算，r 表示压缩率，通常取值为 16。空间注意力模块进一步学习特征图中各空间位置的显著性。该模块首先在通道维度上分别进行平均池化与最大池化操作，提取语义全局与局部信息，随后将二者在通道维度上进行拼接，形成 2 通道的空间描述图。接着通过一个卷积核大小为 3×3 并带有适当填补的卷积层进行空间信息融合，最后通过 Sigmoid 激活函数输出空间注意力图，进而与输入特征进行逐位置加权。

在经过编码网络对水印信号和音频频谱特征进行编码后，输入解码网络进行解码得到嵌入水印的音频频谱特征，如式(4.4)所示：

$$c' = EN(concat(EM_w(sf * w_E), EM_c(c))) \tag{4.4}$$

其中 EN 表示解码器网络，$concat$ 表示连接操作，EM_w 表示水印编码网络，EM_c 表示音频特征编码网络，sf 为嵌入强度因子，默认为 1，w_E 为编码后的水印。对含有水印的音频频谱特征 c' 进行逆短时傅里叶变换，可以将含水印频域特征转换为时域音频信号，如式(4.5)所示：

$$x' = iSTFT(c', p) \tag{4.5}$$

其中 $iSTFT$ 表示逆短时傅里叶变换，c' 表示含水印的音频频谱特征，x' 表示水印音频信号，p 表示嵌入水印前信号的相位特征。得到含水印的音频信号后，输入模拟失真层，并输出失真的含水印音频 $\tilde{x}$，可以表示为式(4.6)：

$$\tilde{x} = Distortion(x') \tag{4.6}$$

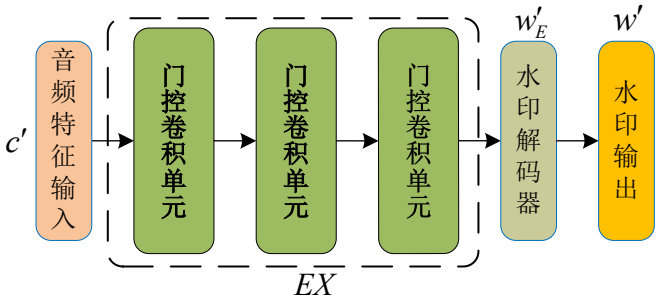


图 4.4 水印提取网络结构图

水印提取网络的目标是从经过失真的含水印音频中完整地提取出水印。本章使用的水印提取网络结构如图 4.4 所示，该结构同样由三个 GCU 模块组成。具体而言，第一个 GCU 模块负责将输入失真音频信号的频谱特征 $\tilde{c}$ 映射到与水印嵌入器编码过程中相同的隐藏维度，将得到的隐藏空间特征输入后续的两个 GCU 模块中进行特征提取得到水印特征 w'_E，如公式(4.7)所示：

$$w'_E = EX(\tilde{c}) \tag{4.7}$$

其中，EX 表示水印提取网络。然后利用与编码相反的线性神经网络对水印特征进行水印解码，得到最终提取的水印 w'，如式(4.8)所示：

$$w' = Linear(Average(w'_E)) \tag{4.8}$$

其中，$Linear$ 表示线性神经网络，$Average$ 表示求均值。鉴别器的目的是促使水印嵌入网络在对音频嵌入水印后水印在音频中不易被人类听觉系统所察觉。鉴别器的主要任务是对原始音频和带有水印的音频进行区分。为此，可以选择用于语音信号分类的神经网络模型作为鉴别器，本章使用的鉴别器由 3 个卷积模块和全连接层组成。鉴别器输入分别为嵌入水印前后音频的短时傅里叶频谱，输出为[0,1]之间的概率。具体而言，前 3 个卷积模块依次由卷积层，实例归一化层和 Leaky ReLU 激活层组成，第一个卷积模块输入通道为 1，输出通道为 16，卷积核大小为 3×3，第二个卷积模块输入通道为 16，输出通道为 32，卷积核大小为 3×3，第三个卷积模块输入通道为 32，输出通道为 64，卷积和大小为 3×3，卷积步长为 1，在卷积计算过程中间进行适当填补。在经过最后的卷积层后进行自适应平均池化，将特征图的宽和高经过池化操作得到 1×1 的特征图，最后再利用全连接层将输入的特征图压缩为[0,1]区间内的结果，表示含水印的音频被判定为原始音频信号的概率。

4.2.3 损失函数设计

对于水印嵌入网络，主要有两个约束条件：将水印嵌入音频中并使得水印在音频中无法被感知，以及欺骗鉴别器使得原始语音信号和嵌入水印的音频信号无法区分。

第一个约束通过优化损失 L_G 实现，L_G 用于衡量原始音频和嵌入水印后的音频之间的误差，使用 MSE 函数计算损失，被定义为式(4.9)：

$$L_G = \mathrm{MSE}(x, x') \tag{4.9}$$

其中，MSE 为均方误差函数，x 表示原始音频信号，x' 表示嵌入水印后的音频信号。通过对式(4.9)进行优化可以最小化原始音频和嵌入水印后音频之间的差异，进而促使水印嵌入网络能生成更逼近于原始音频信号的带水印音频信号。

第二个约束用于衡量鉴别器对嵌入水印后的音频信号分类为未嵌入水印音频信号的成本，通过优化欺骗损失 L_{fool} 进行实现，定义为：

$$L_{fool} = \mathrm{E}[\log(1 - D(c'))] \tag{4.10}$$

其中，E 为交叉熵损失函数，$D(c')$ 表示鉴别器将含水印音频频谱特征 c' 分类为未添加水印信号频谱特征的概率。通过对式(4.10)进行优化可以最小化含水印音频频谱特征 c' 和未添加水印音频频谱特征 c 之间的差异，进而促使水印嵌入网络在嵌入水印时产生更小的扰动，使含水印音频样本和原始音频样本的频谱分布趋于一致。

对于水印提取网络，其约束条件为从含有水印的音频中正确提取出嵌入的水印信号。此约束通过优化水印提取损失 L_w 实现，L_w 用于衡量水印提取网络从含水印语音 x' 中提取到的水印 w' 和目标水印 w 之间的差异，定义为式(4.11)：

$$L_w = \mathrm{MSE}(w', w) \tag{4.11}$$

水印嵌入网络和水印提取网络的优化总损失 L_{total} 由嵌入损失、提取损失和欺骗损失共同加权得到：

$$L_{total} = \alpha L_G + + \beta L_w + \gamma L_{fool} \tag{4.12}$$

其中，α，β，γ 是用来平衡这三种损失的超参数。考虑到鉴别器主要是用于区分原始音频信号 x 与含水印音频 x'，则鉴别器所对应的损失函数 L_D 被定义为：

$$L_D = (\mathrm{E}[\log(1 - D(c))] + \mathrm{E}[\log(D(c'))]) / 2 \tag{4.13}$$

其中，$D(c)$ 表示原始音频信号频谱特征被鉴别器分类为真实样本的概率，$D(c')$ 表示含水印的音频频谱特征被鉴别器分类为真实样本的概率，$D(c)$ 和 $D(c')$ 的取

值范围都在[0,1]内。对式(4.13)对鉴别器进行优化可以使其具有很强的区分原始音频信号 x 和嵌入水印后音频信号 x' 的能力。

4.3 实验结果及分析

4.3.1 实验设置

为了验证所提出方法的有效性，本节同样在 Libri2Mix 数据集上进行实验。其中对数据的处理方式与第三章相同。训练过程的迭代次数设置为 100，损失函数的权重 α，β，γ 分别设置为 10，1，0.01。批量大小设置为 8，水印比特位数默认为 32，强度因子默认为 1，优化器为 Adam，且 $\beta_1 = 0.9$，$\beta_2 = 0.98$，学习率固定为 2×10^{-5}，随机种子统一设为 42。实验的硬件平台为 Intel Core i7 10700 + RTX3090，CPU 主频为 2.9GHz，显存容量为 24GB。软件平台为 Python 3.8 + Pytorch1.12.1 + cuda12.2。

对音频进行处理时，首先将所有音频的采样率统一采样至 16000Hz，在对音频进行短时傅里叶变换时，傅里叶变换的点数设置为 1024，帧移长度设置为 256，窗口长度设为 1024。在神经网络训练时，本章中所使用的模拟失真层与第三章相同。

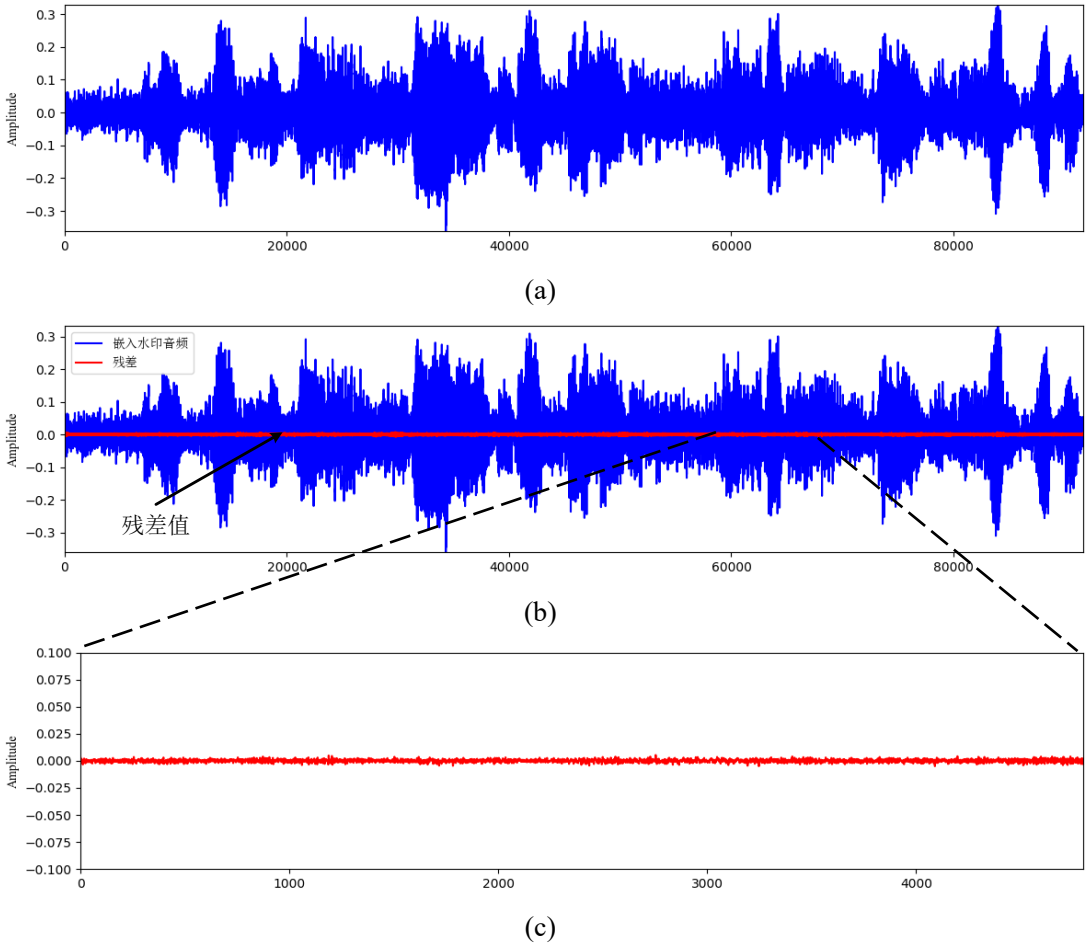
4.3.2 水印隐蔽性分析

由于在实际场景中，音频水印算法主要用于版权保护，所以要求水印嵌入引起的听觉失真应不被察觉，否则嵌入水印的音频可能损失相应的商业价值。因此，需要确保音频在嵌入水印后拥有较高的听觉质量，避免音频应用价值受到破坏。

表 4.1 不同算法下含水印音频的不可感知性评估

方法	SNR(dB)	PESQ	MOS
文献[35]	28.64	4.27	4.56
文献[34]	26.12	3.50	4.38
文献[33]	34.18	4.17	4.73
本章算法	31.56	4.34	4.87

为了测试水印的不可感知性，实验对 Libri2Mix 数据集中测试集的 100 个音频进行水印嵌入，计算水印嵌入前后音频的 SNR 和 PESQ 值。对于主观评价，从中选取 50 个音频样本邀请 10 个受试者进行主观听觉打分，最终得到实验结果如表 4.1 所示。其中文献[35]和文献[33]为基于 STFT 的音频水印算法，文献[34]为基于 DWT 的音频水印算法，实验所用参数与对应论文中的设置保持一致。从中可以观察到，本章所提出算法的平均信噪比可以达到 31.56dB，虽然该算法的平均 SNR 低于文献[33]中提出的算法，但相较于其余算法有一定的提高。此外，该算法在平均 PESQ 得分方面表现出了较高性能，这意味着水印音频在听觉质量上具有出色的表现。在主观评价方面，受试者对大多数样本都给出较高的 MOS 分数，无法察觉到失真。上述实验结果表明本章算法能够获得与当前主流方法相当的听觉质量，具有较高的不可感知性。



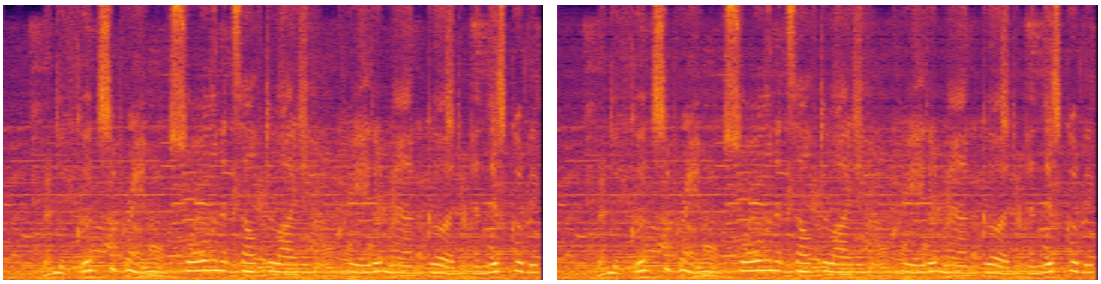
(a)

(b)

(c)

图 4.5 嵌入水印前后音频时域波形图：(a)嵌入水印前, (b)嵌入水印后，红色部分为残差值, (c)放大 10 倍后的局部残差值

图 4.5 展示了从 Libri2Mix 测试集中随机选取的音频样本在嵌入水印前后的时域波形及残差波形图。其中，图 4.5(a)为嵌入水印前，图 4.5(b)为嵌入水印后，图 4.5(c)为前二者残差值变化。从中可以观察到，在时域上，水印算法对原始音频的波形扰动极小，变化难以被感知。从残差波形图中可以看到水印音频与原始音频的残差值的绝对幅度极小，且相对变化均不超过原始音频信号幅度的 1%，且水印带来的噪声在整个时域上分布较为均匀，因此很难被人耳所察觉。图 4.6 展示了嵌入水印前后的频谱可视化结果。从图示结果可看出，嵌入水印后的频谱与原始信号几乎保持一致，能量分布未出现明显波动或异常突变，说明水印嵌入对音频频率成分的影响也极为有限。综合时域和频域两个角度的可视化分析结果，表明了本算法具有良好的不可感知性。



(a) (b)

图 4.6 嵌入水印前后音频频谱对比图：(a)嵌入水印前，(b)嵌入水印后

4.3.3 有效载荷分析

本实验旨在评估所提出音频水印算法在不同嵌入容量下的性能表现，从而确定水印算法的嵌入容量。实验选用 Libri2Mix 测试集中的音频，水印嵌入比特位分别设置为 16 位、32 位、64 位和 128 位。评价指标采用嵌入水印后音频的信噪比以及 PESQ 和水印提取准确率 Acc。除嵌入水印之外，实验的过程中不对音频施加额外的攻击操作。

表 4.2 嵌入不同长度比特水印时各项性能指标

载荷(比特)	SNR(dB)	PESQ	Acc(%)
16	33.53	4.39	100
32	31.56	4.34	100
64	29.85	4.21	99.84
128	27.67	4.04	98.25

实验结果如表 4.2 所示，随着有效载荷的增加，音频的 SNR 呈下降趋势，且水印的提取准确率在嵌入超过 64 比特后也出现了下降趋势，说明嵌入容量与水印系统的不可感知性和鲁棒性之间存在着明显的权衡关系。尽管存在上述现象，使用本算法嵌入 128 比特水印后的音频仍拥有较高的听觉质量和水印提取准确率，表明本章算法能取得良好的有效载荷性能。

4.3.4 水印鲁棒性分析

稳健的音频水印算法必须能够有效应对常见的信号攻击，本节实验通过测试算法面对不同种类信号攻击时的水印提取准确率，对算法的鲁棒性进行了评估。

表 4.3 不同算法在面对常见信号攻击时的表现(%)

攻击类型	文献[35]	文献[34]	文献[33]	本章算法
无攻击	100	100	100	100
Mp3 压缩(64kbps)	100	100	98.00	99.66
Mp3 压缩(128bps)	99.98	100	99.56	99.84
20dB 随机噪声	99.76	99.98	75.86	100
8 比特量化	99.62	96.22	98.65	99.69
回声处理	100	100	99.94	100
低通滤波	90.30	99.97	99.67	99.78
重采样 1	100	96.22	100	100
重采样 2	99.40	97.33	100	100
中值滤波	100	99.98	99.77	100
幅度畸变	100	100	100	100

本节实验选取 Libri2Mix 测试集中的样本进行测试，测试样本采样率为 16kHz，每段音频长度约为 5 秒，内容覆盖不同性别与不同说话风格的语音，具备良好的多样性，嵌入强度因子默认为 1，最终得到的结果如表 4.3 所示。对于 Mp3 压缩攻击，在 64kbps 和 128kbps 两种不同的压缩率下，本章算法提取准确率分别为 99.66%和 99.84%，与文献[35]和文献[34]相当，且都接近 100%。对于 20dB 随机噪声攻击，本章所提出的算法能够达到 100%的提取准确率，相较于其余算法有一定提高；在重采样攻击中，采用了两种重采样方式：第一种为上采样至 32000Hz 再下采样至 16000Hz；第二种为下采样至 8000Hz 再上采样至

16000Hz。对于上述两种重采样攻击，本章所提出算法的水印提取准确率均为100%。面对中值滤波攻击时，本章所提出的算法和文献[35]中的方法都达到 100%的提取准确率，而文献[34]和文献[33]中提出的方法则稍差。对于幅度畸变攻击，幅度畸变的幅度与第三章相同，为放大或缩小 10%，本文与其余方法均对幅度畸变攻击拥有良好鲁棒性。综上所述，与现有算法相比，所提出的算法在面对常见电子攻击时同样具有良好的鲁棒性。

4.3.5 水印在声源分离场景下的性能评估

鲁棒性测试：为了测试本章算法在声源分离场景下的鲁棒性，使用声源分离模型 SepFormer 和 Conv-tasnet 进行实验。实验对 Libri2Mix 数据集上的随机 10 个音频样本使用水印嵌入网络进行水印嵌入，实验中所使用的样本与第 3.3.4 节使用的样本完全一致。对样本嵌入长度为 32 比特的水印，使用声源分离模型将其中混合的音频分离为单个独立的声源。利用水印提取网络对声源分离前的混合音频以及分离后的独立声源音频分别进行水印提取，得到实验结果如表 4.4 所示。

表 4.4 本章算法在不同声源分离模型攻击下的水印提取准确率(%)

样本	水印音频	SepFormer		Conv-tasnet	
		声源 1	声源 2	声源 1	声源 2
样本 1	100	96.88	100	87.50	90.63
样本 2	100	100	90.63	93.75	90.63
样本 3	100	96.88	100	90.63	93.75
样本 4	100	100	100	87.50	87.50
样本 5	100	93.75	100	100	75.00
样本 6	100	96.88	96.88	93.75	100
样本 7	100	100	93.75	90.62	96.88
样本 8	100	93.75	96.88	84.38	81.25
样本 9	100	96.88	96.88	81.25	100
样本 10	100	93.75	96.88	78.13	71.88
平均	100	96.88	97.19	88.75	88.75

从表中结果可看出，未分离的水印音频水印提取准确率为 100%，表明此时的水印结构未被破坏。经过不同声源分离模型处理后，水印提取准确率略有下降，

且在这两种模型的鲁棒性表现存在一定的差异。从表 4.4 可以看出，在 SepFormer 模型分离下，平均提取准确率分别为声源 1 的 96.88%和声源 2 的 97.19%。观察具体样本可发现，样本 2 和样本 7 的声源 2 提取准确率稍低，分别为 90.63%和 93.75%，但提取准确率仍高于 90%，说明分离声源过程虽然会对音频信号造成一定扰动，但仍能够较好地保留音频频谱中的关键信息，保留了水印所依附的特征区域。总体来说，该算法的整体提取效果较为稳定，两个声源的准确率差异较小，说明本章的水印算法具备较强的抗声源分离攻击能力。在 Conv-tasnet 模型分离下的实验结果中，声源 1 的平均提取准确率为 88.63%，而声源 2 降至 88.53%，整体下降了 8%，并且样本之间的提取准确率波动明显更大。样本 10 的声源 1 和声源 2 的准确率分别降至 78.13%和 71.88%，为所有样本中最低。样本 8 的声源 1、声源 2 分别为 84.38%和 81.25%，也远低于平均水平。这一差异反映出本章算法在面对不同声源分离模型时提取准确率存在一定波动。由上述实验可知，本章算法在使用 SepFormer 模型分离声源时达到超过 95%的水印提取准确率，使用 Conv-tasnet 模型时也能达到近 90%的水印提取准确率。因此，相较时域方法，本章所提出频域音频水印算法能够更好地在分离出的声源中保留水印信息，从而实现此类场景下的版权保护目的。

水印透明性：水印的透明性是指嵌入水印后，是否对后续的音频处理产生影响。就本章算法而言，音频在嵌入水印后，使用声源分离模型处理的结果应在听觉质量上与未加入水印时相近，避免引起音频使用者的怀疑。为了探究嵌入水印对声源分离过程的影响，本实验采用 SepFormer 作为声源分离模型，在 Libri2Mix 测试集上进行实验。首先对未嵌入水印的音频样本进行声源分离处理，随后在相应音频中嵌入水印，再次执行声源分离操作。为直观展示水印嵌入对分离效果的影响，从实验中选取了 5 个代表性样本进行可视化分析，结果如图 4.7 所示。其中，图 4.7(a)展示的是未嵌入水印的音频在经过声源分离后的结果，图 4.7(b)则对应嵌入水印后的音频分离结果。图示结果可见，嵌入水印前后的音频波形差异极小，整体结构保持一致，未观察到明显的幅度失真或特定的异常波动。该结果表明，所提出的水印嵌入方法不会对声源分离过程造成可见影响，具有良好的透明性与兼容性。

综上所述，本章提出的基于生成对抗网络的频域音频水印算法相较于时域嵌入能够更好的抵抗声源分离攻击，且嵌入的水印较为隐蔽，不会对分离结果产生明显影响，为实际部署提供了可行性依据，也为后续溯源、鉴权与防伪等安全需求提供了技术支撑。

(a)

(b)

图 4.7 水印嵌入前后声源分离结果对比图：(a)未嵌入水印, (b)嵌入水印

4.3.6 消融实验

强度因子对鲁棒性和不可感知性的影响：为了验证强度因子对算法鲁棒性和隐蔽性的影响，本章针对不同的强度因子进行了实验，实验采用声源分离模型为SepFormer。图 4.8 展示了强度因子对未经过攻击的原始音频的信噪比以及水印提取准确率和分离出的声源水印提取准确率的影响，图 4.8(a)为不同嵌入强度因子对信噪比的影响，图 4.8(b)为不同嵌入强度因子对水印提取准确率影响。从图中可以观察到，在强度因子为 0.1 时，嵌入水印后的音频拥有较高听觉质量，其信噪比达到 36dB 以上。随着嵌入强度因子的增加，水印的不可感知性也随之降低，在强度因子为 1.2 时表现出了最低。而水印提取准确率则随着嵌入强度因子的增大而增大，其中在强度因子为 1.2 时，水印音频和分离后的声源都达到了最高的水印提取准确率。

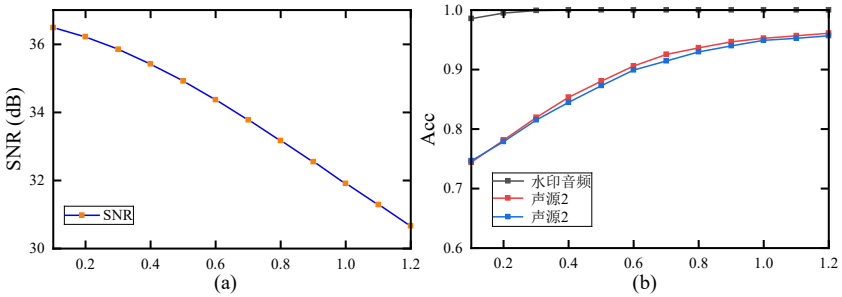


图 4.8 强度因子对性能的影响：(a)不可感知性影响, (b)鲁棒性影响

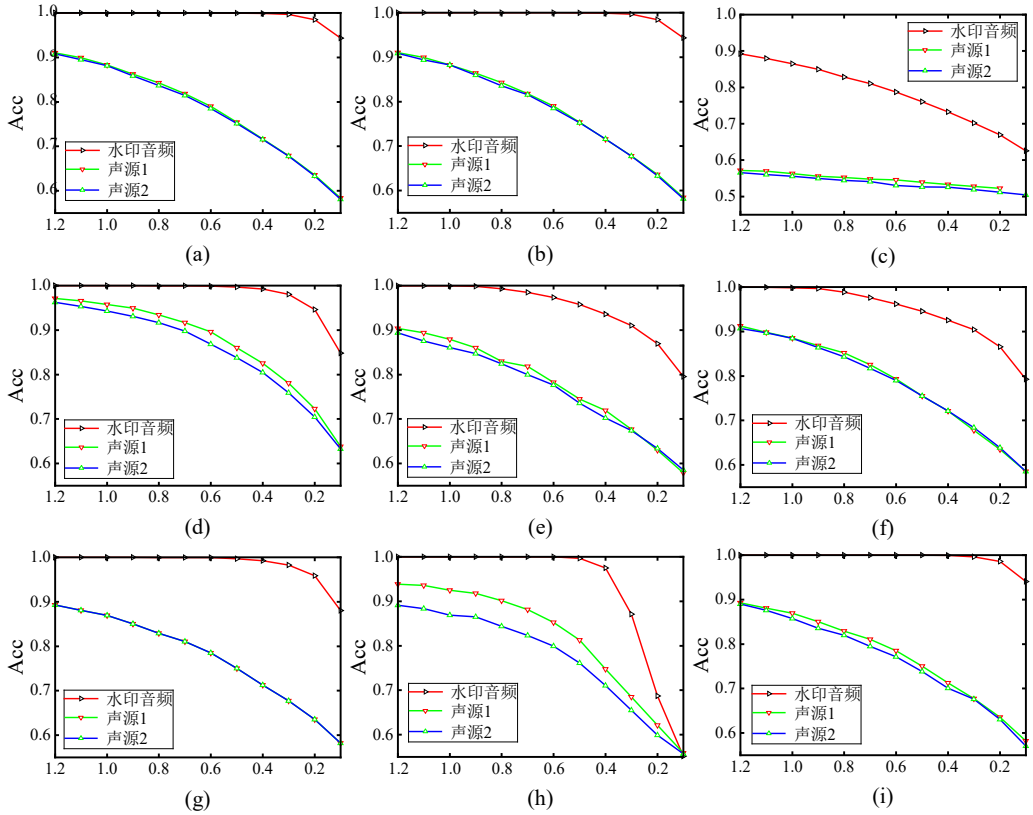


图 4.9 不同攻击下强度因子对水印提取准确率的影响：(a)采样至 32000Hz 再采样回 16000Hz, (b)采样至 8000Hz 再采样回 16000Hz, (c)20dB 随机噪声, (d)幅度畸变(10%), (e) Mp3 压缩(64kbps), (f) Mp3 压缩(128kbps) , (g)中值滤波, (h) 3000Hz 低通滤波, (i) 8 比特量化

图 4.9 展示了混合音频以及分离后的各个音频在面对各种攻击时的水印提取准确率。从图中可以看出，在面对除随机噪声和低通滤波外的攻击时，未分离音频中的水印都具有良好的鲁棒性，即使强度因子为 0.1 时，也能拥有较高的提取准确率。面对低通滤波和随机噪声攻击且强度因子较低时，水印在未分离的音频中表现出了较差的鲁棒性，其中在强度因子为 0.1 时面对两种 Mp3 压缩攻击时提取准确率仅为 80%，而随机噪声攻击后水印提取准确率不足 65%，低通滤波攻击后也不足 60%。

对于分离出的声源 1 和声源 2 中的水印，在强度因子为 1.2 时鲁棒性较高，除低通滤波和白噪声攻击以外，水印提取准确率均能达到 90%。具体来说，对分离出的声源加入 20dB 的白噪声后，即使强度因子为 1.2，从含噪声源的水印提取准确率仍不足 60%，低通滤波攻击时，当强度因子小于 0.6 后，水印提取准确率出现了较大幅度的下降。但从整体上看，分离出的声源信号在面对常见信号攻击仍具有一定的鲁棒性。当强度因子设置为 0.6 时，在不考虑信号攻击的前提下，分离后的声源均能达到 90%的正确率，此时嵌入水印后的音频信噪比也可达到 34dB 以上，表明此时的水印强度设置对于水印的鲁棒性和不可感知性都拥有较高平衡。而在面对常见信号攻击时，除 20dB 白噪声以外，算法也均能达到 70%以上水印提取准确率。因此，本章算法通过引入强度因子提高了算法在水印嵌入幅度上的灵活性，使水印算法在鲁棒性和不可感知性之间取得了较好的权衡。

失真层对于水印抗声源分离能力的影响：为了验证模拟失真层对于水印算法在面对声源分离攻击时的重要性，本章在不同的失真层上进行了不同的水印算法训练，并对使用不同失真层得到的结果在 SepFormer 以及 Conv-tasnet 两个声源分离模型上进行了实验。从表 4.6 可以看出，在失真层加入所有失真种类后，面对两种声源分离攻击算法都有较好的表现。其中在 SepFormer 模型上可达超过 95%的水印提取准确率，而在 Conv-tasnet 声源分离模型上稍差，但也可达到接近 90%的正确率。当在失真层中去掉 RN 和 AD 以后，两种声源分离模型的攻击下的音频水印提取准确率都有一定下降，但仍能保持 80%以上。当去掉低通滤波器以后，水印提取准确率则有较大下降，究其原因，可能是在分离的过程中，高频信息会较大地丢失，导致水印提取准确率下降。因此，对于抵抗声源分离攻击的鲁棒音频水印，良好的模拟失真层是不可缺少的。

表 4.6 不同失真层下水印面对声源分离攻击时提取准确率(%)

失真类型	SepFormer		Conv-tasnet	
	声源 1	声源 2	声源 1	声源 2
LF+RN+AD	96.88	97.19	88.75	88.75
LF+RN	88.38	86.36	81.36	77.45
LF+AD	90.42	88.45	83.56	82.36
RN+AD	78.66	74.44	70.16	72.60

4.4 本章小结

本章针对含有多个声源信号的混合语音音频场景，提出了一种基于生成对抗网络的音频水印算法。首先，介绍了算法的整体框架，包括信号预处理、水印嵌入模块、水印提取模块以及鉴别器的设计，并详细阐述了各模块的具体实现方式。其中，采用频域方式进行水印嵌入是本算法在面对声源分离攻击时具备较强鲁棒性的关键所在。随后，通过不可感知性和鲁棒性两个方面的系统实验对所提算法进行了性能评估，并与相关已有方法进行了对比分析。此外，本章还针对两种主流的声源分离算法设计了水印提取实验，进一步验证算法在实际复杂环境下的有效性。实验结果表明，本算法在保持良好音频感知质量的前提下成功实现了水印的嵌入与提取，表现出较高的不可感知性。同时，在多种典型信号处理攻击下仍能稳定提取水印，展现出较强的鲁棒性。水印算法在声源分离攻击场景下，可以在多个分离声源中恢复水印，且即使分离声源再遭受攻击，水印仍具备较强的抵抗能力。综上所述，本章提出的音频水印算法在数字版权保护与音频内容安全等应用中有较强的综合性能，具有良好的实际应用前景。

第五章　总结与展望

5.1　本文工作总结

声源分离技术用于从混合声源信号中提取出单个独立的声源信号，在人声识别和语音增强等任务中具有重要价值。然而，此类技术也引发了新的安全问题：在应用已有的音频水印算法时，声源分离虽然不会给音频信号带来听觉感知上的失真，但可能影响信号中包含的水印结构，导致水印无法顺利提取。针对水印算法在声源分离场景下鲁棒性不足的问题，本文对声源分离前后的音频信号进行了深入的研究和分析，对声源分离过程中存在的失真进行了建模，据此分别在时域和频域两个角度上，提出了两种抗声源分离攻击的音频水印方案。本文主要工作总结如下：

1）针对音频水印在面对声源分离攻击时无法准确提取的问题，本文提出了一种基于可逆神经网络的时域音频水印框架。在嵌入水印前，对音频信号进行分帧，将水印信号编码至与分帧后音频信号相同的维度，并与分帧音频同时输入至可逆神经网络进行训练。为了使算法能够抵抗声源分离攻击，训练过程中对此攻击进行建模并设计了失真层，以模拟声源分离过程中出现的失真。由于水印嵌入在音频的低频区域，为了使水印具备更好的不可感知性，引入低频损失约束低频区域的水印嵌入强度。实验结果表明，相比现有方法，该方法使得音频在嵌入水印后拥有较高不可感知性的同时，面对常见电子攻击时也具有较强鲁棒性。此外，该方法能在分离后的多个声源中提取出水印，表明其能够在一定程度上抵抗声源分离攻击。

2）针对在时域中嵌入水印鲁棒性不足的问题，本文提出了一种基于生成对抗网络的频域音频水印算法。基于对音频信号频谱的分析，本文发现声源分离前后频谱的低频区域变化较小。因此，该算法选择在音频低频区域的频谱特征中嵌入水印。具体而言，将水印信息编码至与信号频谱特征相同的维度，并与信号频谱一同输入至生成器中，通过嵌入强度因子对水印幅度进行约束，进而得到高保真的含水印音频。随后利用解码器从含水印音频的频谱特征中恢复出水印。鉴别器用于引入对抗训练，促使生成器在嵌入过程中学习鲁棒且隐蔽的嵌入策略的同

时，提高鉴别器自身对含水印频谱的判别能力。实验结果表明，所提出的频域音频水印方法在实现高不可感知性的同时，能够抵抗常见的音频信号处理操作。在面对不同的声源分离攻击算法时，该方法相较时域方法展现出更高的水印提取准确率。此外，在不同嵌入强度因子和模拟失真层上的消融实验充分表明，该方法有效实现了不可感知性与鲁棒性之间的平衡。

5.2 未来工作展望

本文旨在保护音频水印不受声源分离攻击的影响，深入分析了声源分离过程中带来的失真，并从时域和频域角度分别提出了两种不同的水印算法，提高了音频水印算法的通用性。在未来研究中，为进一步提升音频水印算法面对声源分离攻击时的性能，可以从以下方向进行发展：

1）在时域和频域协同嵌入水印，使水印相互补偿、协同抵御不同类型的分离失真。同时，可借助自适应权重或注意力机制，根据局部信噪比与听觉掩蔽阈值动态调整各域嵌入强度，形成灵活的鲁棒与不可感知平衡。在此基础上，构建端到端联合训练框架，让网络自动学习最优的多域信息分配与融合方式，提升水印在复杂声源分离场景中的生存能力。

2）引入声源分离对抗训练。在神经网络训练过程中，将水印嵌入网络、声源分离网络与水印提取网络组成统一的框架，采用对抗式目标函数对各个网络进行联合优化。通过动态博弈迫使水印系统面对较强分离攻击时仍能保持稳定提取准确率，从而获得更高的鲁棒性。

3）可在水印框架中引入“失真补偿”机制。通过在模拟失真层和提取网络之间插入可学习的补偿模块，对声源分离造成的幅度畸变、相位扰动和时序错位进行反向修复；同时采用自适应估计策略，根据分离残差实时调整补偿参数。这样不仅可提高水印在复杂失真下的提取率，还能在保持不可感知性的前提下进一步增大嵌入容量。

参考文献

[1] 韩冰. 数字广播技术的特点及其应用[J]. 电声技术, 2025, 49(03): 106-108.

[2] TURCHET L, FAZEKAS G, LAGRANGE M, et al. The internet of audio things: State of the art, vision, and challenges[J]. IEEE Internet of Things journal, 2020, 7(10): 10233-10249.

[3] 邢巧莲. 新质生产力背景下人工智能语音技术在高职英语听力教学中的创新应用研究[J]. 中国多媒体与网络教学学报(中旬刊), 2024, (10): 5-8.

[4] KIETZMANN J, LEE L W, MCCARTHY I P, KIETZMANN T C. Deepfakes: Trick or treat?[J]. Business Horizons, 2020, 63(2): 135-146.

[5] KORSHUNOV P, MARCEL S. Deepfakes: A new threat to face recognition? Assessment and detection[J]. arXiv Preprint arXiv:181208685, 2018.

[6] PATIL A, HIRAN D, SHELKE R. Ha2m-opm: Hybrid adaptive amplitude modulation and optiphase modulation technique for digital audio watermarking[J]. Circuits, Systems, and Signal Processing, 2025: 1-36.

[7] HUA G, HUANG J, SHI Y Q, et al. Twenty years of digital audio watermarking—a comprehensive review[J]. Signal processing, 2016, 128: 222-242.

[8] WEN S, ZHANG Q, HU T, LI J. Robust audio watermarking against manipulation attacks based on deep learning[J]. IEEE Signal Processing Letters, 2025, 32: 126-130.

[9] ABBASI A T, MIAO F, ISLAM M S. A secure and robust audio watermarking scheme using secret sharing in the transform-domain[J]. Circuits, Systems, and Signal Processing, 2025, 44(2): 1274-1307.

[10] PARSONS T W. Separation of speech from interfering speech by means of harmonic selection[J]. The Journal of the Acoustical Society of America, 1976, 60(4): 911-918.

[11] VIRTANEN T. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(3): 1066-1074.

[12] ZHANG X, TANG J, CAO H, et al. Cascaded speech separation denoising and dereverberation using attention and tcn-wpe networks for speech devices[J]. IEEE Internet of Things Journal, 2024, 11(10): 18047-18058.

[13] HINTON G, DENG L, YU D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups[J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.

[14] GRAVES A, MOHAMED A-R, HINTON G. Speech recognition with deep recurrent neural networks[C]//2013 IEEE international Conference on Acoustics, Speech and Signal Processing, May 5-14, 2013, New York, America. Piscataway: IEEE, 2013: 6645-6649.

[15] 马润泽. 基于声源分离的声音事件检测算法设计与实现[D]. 上海: 上海大学, 2022.

[16] XU Y, DU J, DAI L-R, LEE C-H. A regression approach to speech enhancement based on deep neural networks[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 23(1): 7-19.

[17] SERBAN I, SORDONI A, LOWE R, et al. A hierarchical latent variable encoder-decoder model for generating dialogues[C]//Proceedings of the AAAI Conference on Artificial Intelligence, February 4-9, 2017, California, USA. Menlo Park: AAAI, 2017.

[18] COX I J, MILLER M L, BLOOM J A, et al. Chapter 1 - introduction[M]. The morgan kaufmann series in multimedia information and system: Digital watermarking and steganography (second edition). Burlington; Morgan Kaufmann. 2008: 1-13.

[19] 杨艳聪, 王锐, 秦兴红, 刘正辉. 抗重录音的溯源追踪音频水印算法[J]. 计算机应用与软件, 2024, 41(08): 376-381.

[20] FASTL H, ZWICKER E. Stimuli and procedures[M]//FASTL H, ZWICKER E. Psychoacoustics: Facts and models. Berlin, Heidelberg; Springer Berlin Heidelberg. 2007: 1-15.

[21] GUPTA S, GOYAL A, BHUSHAN B. Information hiding using least significant bit steganography and cryptography[J]. International Journal of Modern Education and Computer Science, 2012, 4(6): 27.

[22] BENDER W, GRUHL D, MORIMOTO N, LU A. Techniques for data hiding[J]. IBM systems journal, 1996, 35(3.4): 313-336.

[23] CVEJIC N, SEPPANEN T. Increasing the capacity of lsb-based audio steganography[C]//2002 IEEE Workshop on Multimedia Signal Processing, December 9-11, 2002, St. Thomas, VI, USA. Piscataway: IEEE, 2002: 336-338.

[24] SUGATHAN S. An improved lsb embedding technique for image steganography[C]//2016 2nd international Conference on Applied and Theoretical Computing and Communication Technology July 21-23, 2016, Bangalore, India. Piscataway: IEEE, 2016: 609-612.

[25] JAYARAM P, RANGANATHA H, ANUPAMA H. Information hiding using audio steganography–a survey[J]. The International Journal of Multimedia & Its Applications (IJMA) Vol, 2011, 3: 86-96.

[26] ERFANI Y, SIAHPOUSH S. Robust audio watermarking using improved ts echo hiding[J]. Digital Signal Processing, 2009, 19(5): 809-814.

[27] WANG X-Y, ZHAO H. A novel synchronization invariant audio watermarking scheme based on dwt and dct[J]. IEEE Transactions on Signal Processing, 2006, 54(12): 4835-4840.

[28] LEI B, SOON Y, TAN E-L. Robust svd-based audio watermarking scheme with differential evolution optimization[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2013, 21(11): 2368-2378.

[29] HU H T, WU S T, LEE T T. Fft-based dual-mode blind watermarking for hiding binary logos and color images in audio[J]. IEEE Access, 2023, 11: 37612-37622.

[30] LI P, ZHANG X, XIAO J, WANG J. Ideaw: Robust neural audio watermarking with invertible dual-embedding[J]. arXiv Preprint arXiv:240919627, 2024.

[31] PAVLOVIĆ K, KOVAČEVIĆ S, DJUROVIĆ I, WOJCIECHOWSKI A. Robust speech watermarking by a jointly trained embedder and detector using a dnn[J]. Digital Signal Processing, 2022, 122: 103381.

[32] ROMAN R S, FERNANDEZ P, DéFOSSEZ A, et al. Proactive detection of voice cloning with localized watermarking[J]. arXiv Preprint arXiv:240117264, 2024.

[33] CHEN G, WU Y, LIU S, et al. Wavmark: Watermarking for audio generation[J]. arXiv Preprint arXiv:230812770, 2023.

[34] LIU C, ZHANG J, FANG H, et al. Dear: A deep-learning-based audio re-recording resilient watermarking[C]//Proceedings of the AAAI Conference on Artificial Intelligence, February 7–14, 2023, Washington, USA. Menlo Park: AAAI, 2023: 13201-13209.

[35] LIU C, ZHANG J, ZHANG T, et al. Detecting voice cloning attacks via timbre watermarking[J]. arXiv Preprint arXiv:231203410, 2023.

[36] HYVäRINEN A, OJA E. Independent component analysis: Algorithms and applications[J]. Neural Networks, 2000, 13(4-5): 411-430.

[37] LEE D D, SEUNG H S. Learning the parts of objects by non-negative matrix factorization[J]. Nature, 1999, 401(6755): 788-791.

[38] OZEROV A, FéVOTTE C, BLOUET R, DURRIEU J-L. Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation[C]//2011 IEEE International Conference on Acoustics, Speech and Signal Processing Proceedings, May 22-27, 2011, Prague, Czech. Piscataway: IEEE, 2011: 257-260.

[39] LEE D, SEUNG H S. Algorithms for non-negative matrix factorization[J]. Advances in Neural Information Processing Systems, 2000, 13.

[40] HERSHEY J R, CHEN Z, LE ROUX J, WATANABE S. Deep clustering: Discriminative embeddings for segmentation and separation[C]//2016 IEEE international Conference on Acoustics, Speech and Signal Processing, March 20-25, 2016, Shanghai, China. Piscataway: IEEE, 2016: 31-35.

[41] LUO Y, MESGARANI N. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(8): 1256-1266.

[42] LEA C, FLYNN M D, VIDAL R, et al. Temporal convolutional networks for action segmentation and detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, July 22-25, 2017, Honolulu, USA. Piscataway: IEEE, 2017: 156-165.

[43] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, December 4-9, 2017, California, USA. Red Hook, NY: Curran Associates Inc, 2017: 6000–6010.

[44] SUBAKAN C, RAVANELLI M, CORNELL S, et al. Attention is all you need in speech separation[C]//2021 IEEE International Conference on Acoustics, Speech and Signal Processing Proceedings, June 06-11, 2021, Toronto, Canada. Piscataway: IEEE, 2021: 21-25.

[45] HOPFIELD J J. Neural networks and physical systems with emergent collective computational abilities[J]. Proceedings of the National Academy of Sciences, 1982, 79(8): 2554-2558.

[46] CHEN J, MAO Q, LIU D. Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation[J]. arXiv Preprint arXiv:200713975, 2020.

[47] JING J, DENG X, XU M, et al. Hinet: Deep image hiding by invertible network[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, October 10-17, 2021, Montreal, Canada. Piscataway: IEEE, 2021: 4733-4742.

[48] RAMANA A, PARAYITAM L, PALA M S. Investigation of automatic speech recognition performance and mean opinion scores for different standard speech and audio codecs[J]. IETE Journal of Research, 2012, 58(2): 121-129.

[49] RIX A W, BEERENDS J G, HOLLIER M P, HEKSTRA A P. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs[C]//2001 IEEE international Conference on Acoustics, Speech, and Signal Processing Proceedings, May 07-11, 2001, Salt Lake City, USA. Piscataway: IEEE, 2001: 749-752.

[50] MOULIN P, O'SULLIVAN J A. Information-theoretic analysis of watermarking[C]//2000 IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings, June 05-09, 2000, Istanbul, Turkey. Piscataway: IEEE, 2000: 3630-3633.

[51] 李锵, 陈德昱, 关欣. 时间及通道双维序列注意力音乐声源分离方法[J]. 声学学报, 2023, 48(03): 588-598.

[52] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[53] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-

Assisted Intervention–MICCAI 2015: 18th international conference, October 5-9, 2015, Munich, Germany. Berlin: Springer, 2015: 234-241.

[54] VINCENT E, GRIBONVAL R, FéVOTTE C. Performance measurement in blind audio source separation[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2006, 14(4): 1462-1469.

[55] MITIANOUDIS N, DAVIES M E. Audio source separation: Solutions and problems[J]. International Journal of Adaptive Control and Signal Processing, 2004, 18(3): 299-314.

[56] DURAK L, ARIKAN O. Short-time fourier transform: Two fundamental properties and an optimal implementation[J]. IEEE Transactions on Signal Processing, 2003, 51(5): 1231-1242.

[57] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[58] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1, December 3-8, 2012, Lake Tahoe, Nevada. Red Hook, NY: Curran Associates Inc, 2012: 1097–1105.

[59] LECUN Y, BOSER B, DENKER J S, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural Computation, 1989, 1(4): 541-551.

[60] LECUN Y, BOTTOU L, BENGIO Y, HAFFNER P. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[61] HE K, ZHANG X, REN S, SUN J. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, USA. Piscataway: IEEE, 2016: 770-778.

[62] HUANG G, LIU Z, VAN DER MAATEN L, WEINBERGER K Q. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, July 22-25, 2017, Honolulu, USA. Piscataway: IEEE, 2017: 4700-4708.

[63] DINH L, KRUEGER D, BENGIO Y. Nice: Non-linear independent components estimation[J]. arXiv Preprint arXiv:14108516, 2014.

[64] ARDIZZONE L, KRUSE J, WIRKERT S, et al. Analyzing inverse problems with invertible neural networks[J]. arXiv Preprint arXiv:180804730, 2018.

[65] JADERBERG M, SIMONYAN K, ZISSERMAN A, KAVUKCUOGLU K. Spatial transformer networks[C]//Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 2, December 7-12, 2015, Montreal, Canada. Cambridge, MA: MIT Press, 2015: 2017–2025.

[66] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, June 18-22, 2018, Utah, USA. Piscataway: IEEE, 2018: 7132-7141.

[67] WOO S, PARK J, LEE J-Y, KWEON I S. Cbam: Convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision September 8-14, 2018, Munich, Germany. Verlag: Springer, 2018: 3-19.

[68] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[69] SWANSON M D, ZHU B, TEWFIK A H, BONEY L. Robust audio watermarking using perceptual masking[J]. Signal processing, 1998, 66(3): 337-355.

[70] LEE S-K, HO Y-S. Digital audio watermarking in the cepstrum domain[J]. IEEE Transactions on Consumer Electronics, 2000, 46(3): 744-750.

[71] KONG J, KIM J, BAE J. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems, December 6-12, 2020, Vancouver, BC, Canada. Red Hook, NY: Curran Associates Inc, 2020: 17022--17033.

[72] KIM J, KONG J, SON J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech[C]//International Conference on Machine Learning, July 18-24, 2021, Virtual. New York: PMLR, 2021: 5530-5540.

[73] EMIYA V, VINCENT E, HARLANDER N, HOHMANN V. Subjective and objective quality assessment of audio source separation[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(7): 2046-2057.

[74] HORITA A, NAKAYAMA K, HIRANO A, DEJIMA Y. Analysis of signal separation and signal distortion in feedforward and feedback blind source separation based on source spectra[C]//Proceedings 2005 IEEE International Joint Conference on Neural Networks, 2005, July 31- August 04, 2005, Montreal, Canada. Piscataway: IEEE, 2005: 1257-1262.

[75] COSENTINO J, PARIENTE M, CORNELL S, et al. Librimix: An open-source dataset for generalizable speech separation[J]. arXiv Preprint arXiv:200511262, 2020.

[76] PANAYOTOV V, CHEN G, POVEY D, KHUDANPUR S. Librispeech: An asr corpus based on public domain audio books[C]//2015 IEEE international Conference on Acoustics, Speech and Signal Processing Proceedings, April 19-24, 2015, Brisbane, Australia. Piscataway: IEEE, 2015: 5206-5210.

[77] RAVANELLI M, PARCOLLET T, PLANTINGA P, et al. Speechbrain: A general-purpose speech toolkit[J]. arXiv Preprint arXiv:210604624, 2021.

[78] TRAMèR F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: Attacks and defenses[J]. arXiv Preprint arXiv:170507204, 2017.

[79] DAUPHIN Y N, FAN A, AULI M, GRANGIER D. Language modeling with gated convolutional networks[C]//International Conference on Machine Learning, August 6-11, 2017, Sydney, Australia. New York: PMLR, 2017: 933-941.

[80] GEHRING J, AULI M, GRANGIER D, et al. Convolutional sequence to sequence learning[C]//International Conference on Machine Learning, August 6-11, 2017, Sydney, Australia. New York: PMLR, 2017: 1243-1252.

[81] VAN DEN OORD A, DIELEMAN S, ZEN H, et al. Wavenet: A generative model for raw audio[J]. arXiv Preprint arXiv:160903499, 2016, 12.

攻读硕士学位期间取得的研究成果

[1] Shi J, Yang Z, Li L, et al. Multi-user watermarking based on paralleled invertible neural networks[J]. Signal, Image and Video Processing, 2025, 19(6): 1-12.（第一作者，SCI，WOS:001469991500001）

致　谢

栀子花开，又到毕业的季节。写完这篇毕业论文，宣告着我的研究生生涯即将落幕，也宣布我的学生生涯也到此告一段落。非常有幸能在研究生阶段加入人工智能安全实验室，在这里我学习到了许多信息安全方面的知识，更有幸认识了一些高学术水平的老师。

首先，我要感谢我的导师吴汉舟老师，在硕士阶段的三年学习和生活中，吴老师都给予我悉心指导和耐心帮助。从课题选题到论文撰写，吴老师严谨的治学态度、深厚的专业知识和高尚的人格魅力，始终给予我深刻的影响和莫大的鼓励。同时，我还要特别感谢李莉老师。在我的科研过程中，李老师给予了我极大的帮助与指导。她严谨的学术态度和细致入微的建议让我在研究中少走了很多弯路，也极大地提升了我的科研能力。在此谨向李莉老师表示诚挚的感谢。

我还要感谢实验室的同门以及师兄师姐和师弟师妹们。感谢他们在科研道路上和生活方面给我提供的帮助，感谢他们能在我有疑惑的时候为我答疑，感谢有了他们，我的研究生三年生活变得丰富多彩，衷心祝愿他们的未来一切顺利。跟他们在实验室一起奋斗的时光，对于我的人生来说是一笔宝贵的财富。

此外，我还要感谢我的家人和朋友们。在整个研究生生涯中，他们都是我最坚强的后盾，有了他们我才有底气去挑战。无论我遇到什么挫折与困难，都始终给予我最大的鼓励和帮助，让我能全身心的投入在学术研究中。

最后，感谢百忙之中抽出时间评阅论文的专家老师们，您们的建议和意见对我的学术研究和人生规划都有很大的帮助。在未来的道路上，我会一如既往地严谨、努力，争取取得更多的成就。

作者署名：史景辉

完成地点：上海大学

2025 年 5 月 15 日