中图分类号：　　单位代号：10280
密　　级：　　学　　号：21721329

上海大学

硕士学位论文

SHANGHAI UNIVERSITY
MASTER'S DISSERTATION

题目	高保真的生成式模型水印技术研究

作　　者：张力
学科专业：信号与信息处理
导　　师：张新鹏
完成日期：2024 年 5 月

姓　　名：张力　　学号：21721329

论文题目：高保真的生成式模型水印技术研究

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/博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：张力　　学号：21721329

论文题目：高保真的生成式模型水印技术研究

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

上海大学工学硕士学位论文

高保真的生成式模型水印技术研究

作　　者：张力

学科专业：信号与信息处理

导　　师：张新鹏

上海大学通信与信息工程学院

二〇二四年五月

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

Research on Generative Model Watermarking Techniques with High Fidelity

Candidate： **Li Zhang**

Major： **Signal and Information Processing**

Supervisor： **Xinpeng Zhang**

School of Communication and Information Engineering

Shanghai University

May, 2024

摘 要

模型水印技术是一种用于保护和验证深度学习模型知识产权的技术，它对防止知识产权侵权及维护模型拥有者的权益具有重要的意义。近年来，一些研究已经将模型水印技术应用于图像生成式网络，以保护这些模型的知识产权。虽然现有的生成式模型水印技术有效地实现了版权保护，但它们容易引起高频伪影问题。这些高频伪影不仅降低了生成式模型水印的保真度，而且损害了水印系统的隐蔽性和安全性。为了解决这一问题，本文研究消除高频伪影的高保真生成式模型水印技术。主要工作如下：

1) 针对生成式模型水印中水印嵌入过程所导致的高频伪影问题，本文提出一种高保真的生成式模型水印框架，通过优化输出图像中水印的分布来抑制高频伪影。具体而言，该方法利用小波频域分离层来获取输出图像的低频成分，然后联合训练水印嵌入网络和水印提取网络，从而有效地将水印嵌入到图像的低频区域。实验结果表明，与相关方法相比，本方法有效地抑制了高频伪影，并实现了良好的保真度和隐蔽性。

2) 上述研究通过优化水印在输出图像中的分布来抑制高频伪影，但却未考虑嵌入网络的固有结构可能导致的高频伪影问题。为了解决这一问题，本文提出一种高保真的生成式模型水印框架。具体而言，本方法没有使用嵌入网络来实现水印嵌入，从而避免了嵌入网络所引起的高频伪影问题；频域扰动将作为水印被直接添加到输出图像的低频成分中，从而在完成水印嵌入的同时优化水印在输出图像中的分布。实验结果表明，与相关方法相比，本方法有效地消除了高频伪影，并显著提高了保真度和隐蔽性。

关键词：模型水印；生成式模型；高保真；高频伪影；隐蔽性

ABSTRACT

Model watermarking is a technique used to protect and verify the intellectual property of deep learning models, which is important for preventing intellectual property infringement and maintaining the rights of model owners. In recent years, several research have applied model watermarking techniques to image generative networks to protect them. Although existing generative model watermarking techniques are effective in achieving copyright protection, but they are prone to cause high-frequency artifacts problems. These high-frequency artifacts not only reduce the fidelity of generative model watermarking, but also compromise the invisibility and security of the watermarking system. To solve this problem, this thesis investigates the high-fidelity generative model watermarking technique that eliminates high-frequency artifacts. The main work is as follows:

1) To address the high-frequency artifacts caused by the watermark embedding process, this thesis proposes a high-fidelity generative model watermarking framework that suppresses high-frequency artifacts by optimizing the watermark distribution in the output image. Specifically, the method utilizes a wavelet frequency separation layer to obtain the low-frequency components of the output image, and then co-trains the watermark embedding network and the watermark extraction network to effectively embed the watermark into the low-frequency region of the image. The experimental results show that the proposed method effectively suppresses high-frequency artifacts and achieves good fidelity and invisibility compared with related methods.

2) The above studies suppress high-frequency artifacts by optimizing the distribution of the watermark in the output image, but without considering the high-frequency artifacts that may be caused by the inherent structure of the embedded network. To solve this problem, a high-fidelity generative model watermarking framework is proposed in this thesis. Specifically, this method does not use the embedding network

to achieve watermark embedding, thus avoiding the high-frequency artifacts caused by the embedding network; the frequency-domain perturbation will be added directly to the low-frequency components of the output image as a watermark, thus optimizing the distribution of the watermark in the output image while completing the watermark embedding. The experimental results show that the present method effectively eliminates high-frequency artifacts and significantly improves fidelity and invisibility compared with related methods.

Keywords: Model watermarking; Generative model; High fidelity; High-frequency artifacts; Invisibility

目 录

摘 要 ..... **III**

ABSTRACT ..... **IV**

第一章 绪论 **1**

1.1 课题研究背景与意义 1

1.2 国内外研究现状 3

1.2.1 白盒水印 5

1.2.2 黑盒水印 7

1.2.3 无盒水印 9

1.3 本文研究内容和结构安排 11

1.3.1 研究内容 11

1.3.2 结构安排 12

1.4 本章小结 14

第二章 相关技术基础 **15**

2.1 生成式模型水印 15

2.1.1 生成式模型水印简介 15

2.1.2 生成式模型水印的评价指标 16

2.1.3 生成式模型水印所面临的问题 18

2.2 离散余弦变换 20

2.3 高频伪影的成因分析 21

2.4 本章小结 22

第三章 基于小波变换的生成式模型水印 **23**

3.1 研究动机 23

3.2 方案设计 25

3.2.1 理论分析 25

3.2.2 框架概述 26

3.2.3 结构设计 28

3.2.4 损失函数.....29
3.3 实验结果与分析.....31
3.3.1 实验设置.....31
3.3.2 定性和定量实验结果.....32
3.3.3 对预处理攻击的鲁棒性.....35
3.3.4 对高频信息移除攻击的鲁棒性.....38
3.3.5 与现有方法比较.....40
3.4 本章小结.....40
第四章 基于频域扰动的生成式模型水印.....42
4.1 研究动机.....42
4.2 方案设计.....43
4.2.1 框架概述.....43
4.2.2 结构设计.....44
4.2.3 损失函数.....45
4.3 实验结果与分析.....47
4.3.1 实验设置.....47
4.3.2 定性和定量实验结果.....48
4.3.3 对预处理攻击的鲁棒性.....51
4.3.4 与传统图像水印方法的比较.....55
4.3.5 过拟合问题分析.....56
4.4 本章小结.....57
第五章 总结与展望.....59
5.1 总结.....59
5.2 展望.....60
参考文献.....61
攻读硕士学位期间取得的研究成果.....70
致 谢.....71

第一章 绪论

1.1 课题研究背景与意义

随着人工智能（Artificial Intelligence, AI）技术的迅猛发展，深度学习[1]作为其重要组成部分，推动了 AI 技术在多个领域的广泛应用，如计算机视觉[2]、语音识别[3]、自然语言处理[4]、自动驾驶[5]、医疗影像分析[6]等。从推动 AI 艺术创作的 DALL-E 2[7]和 Stable Diffusion[8]，再到以 ChatGPT[9]为代表的接近人类水平的对话机器人，更为先进的 AI 技术不断涌现，为蓬勃发展的 AI 领域又增添了新的活力。同时，人工智能生成内容（Artificial Intelligence Generated Content, AIGC）已成为科技领域最引人注目的进展之一。从文本、图像到音乐和视频，AIGC 正在改变人们创造和消费内容的方式。

AI 技术所具备的强大能力，能够智能地处理海量数据，实现内容创作与智能推理。这一能力促进了各行业的效率提升和创新突破，从而产生了显著的经济价值和社会价值。基于此，国内外各大公司纷纷投身于 AI 技术，将其纳入核心发展战略，以提高组织效率、产品质量和收益。例如，谷歌、微软、阿里和腾讯等都在大力投资 AI 模型的开发。构建一款性能卓越的 AI 模型通常需要大量资源。在模型训练前，需要获取足量的数据，这个过程涉及到数据的收集、清理、处理和存储等多个环节。此外，深度模型的网络结构和训练方法的设计也依赖于专业人才。最终，训练这些模型通常还需要大规模的计算资源。以 OpenAI 公司推出的 ChatGPT-3 为例，该模型包含 1750 亿参数，其训练过程涉及上万块图形处理器（Graphics Processing Unit, GPU），训练成本接近 460 万美元。而 OpenAI 后续更新的 ChatGPT-4 更是将模型参数增加十倍以上，需要在大约 25000 个 A100 GPU 上训练 90 多天。如果以 OpenAI 云计算的成本为 1 美元/每 A100 小时来计算，仅仅这次训练的成本就约为 6300 万美元[10-11]。

随着 AI 技术能力的不断提升以及各大公司对 AI 模型的持续投入，这些 AI 模型和 AIGC 的价值日益上升。因此，对于模型的所有者来说，保护这些具有高价值的 AI 模型和 AIGC 的知识产权，并维护其核心利益，显得尤为关键。然而，AI 模型很容易被其他人窃取、篡改、贩卖或分发。例如，攻击者可能通过窃取训练数据或模型输出的数据来训练出功能相似的窃取模型；攻击者也可能通过微调、剪枝、蒸馏等攻击手段来获取模型；攻击者也可能通过发现后门水印的异常触发条件来破坏水印。另一方面，AIGC 也可能面临以上的侵权行为，攻击者可能直接窃取 AIGC，并声称对其的所有权。最终，攻击者可以通过售卖非法获取的深度模型或 AIGC 来牟利。近期，AI 领域的侵权案例层出不穷[13]，引起了业界的广泛关注。例如，2018 年，腾讯公司针对网贷之家提起的 AI 写作侵权诉讼，成为国内深度学习模型版权争议的标志性事件。进一步地，在 2023 年，多位艺术家对小红书平台的 AI 工具涉嫌数据窃取的集体诉讼，又开创了国内 AI 模型训练数据侵权的先例。这些具有里程碑意义的事件不仅凸显了深度学习模型知识产权保护的紧迫性，也反映出越来越多的企业和组织正成为此类纠纷的参与者。侵权行为的频繁发生，对人工智能企业的核心利益构成了前所未有的挑战。这不仅削弱了企业的创新动力和市场竞争力，而且严重破坏了人工智能产业的健康生态，对产业的可持续发展构成了显著威胁。鉴于此，强化深度学习模型的知识产权保护，已成为推动人工智能产业高质量发展的关键要素。

近年来，国家层面对 AI 技术的安全应用给予了前所未有的重视，并逐步构建和完善了相应的政策法规体系。在国家科技发展战略的宏观布局中，强化对深度学习模型等核心技术知识产权的保护变得尤为关键。2017 年，国务院出台了《新一代人工智能发展规划》，该规划明确指出了建立统一的人工智能技术标准和知识产权保护体系的紧迫性和重要性。随后，在 2020 年，中国信息通信研究院颁布了《人工智能安全框架》，其中提出了“模型水印”概念，为 AI 知识产权保护开辟了新的路径。进一步地，2023 年，国家网信办联合其他七个部门共同发布了《生成式人工智能服务管理暂行办法》，旨在规范和促进生成式人工智能的健康发展。这些政策文件的相继出台，不仅体现了国家对 AI 安全发展的高度

关注，也标志着深度学习模型知识产权保护已经上升为国家创新发展战略的核心需求。

综合上述分析，保护深度学习模型的知识产权已不仅是人工智能产业高质量发展的关键驱动力，更是国家创新发展战略中不可或缺的重要组成部分。因此，开发和实施深度模型知识产权保护的创新解决方案，以及推动相关技术的发展刻不容缓。因此，深度学习模型版权保护的相关研究已成为学术界和工业界的研究热点，该领域的研究可以看作是传统版权保护在全新载体（即深度模型）上的延伸。传统的多媒体版权保护研究始于上个世纪 50 年代，至今已经有 70 多年的历史。数字水印是实现版权保护的主要手段之一，主要应用于保护图像、音频、视频等多媒体内容。借鉴数字水印技术的理念，模型水印技术[10-13]应运而生。该技术的核心目标是在深度学习模型中巧妙地嵌入特定的水印信息。这样，模型的所有者就能够在需要时从模型中准确提取这些预先嵌入的水印信息，从而有效地进行知识产权的认证和保护。2017 年，Uchida 等人[14]提出了第一个模型水印算法，随即引发了国内外研究人员的广泛关注。经过多年的发展，模型水印技术已经可以在许多实际场景中应用，但模型水印技术仍然存在着许多待解决的难题。为了迎合深度模型知识产权保护的紧迫需求，推动模型水印技术的发展刻不容缓。

1.2 国内外研究现状

模型水印技术是目前主流的深度模型知识产权保护方法，一经提出，立即引起了国内外众多研究人员的关注。根据版权验证阶段所需的不同条件，模型水印方法可以大致分为三类：白盒水印、黑盒水印、无盒水印。如表 1.1 展示了不同模型水印方法在版权验证阶段需要满足的条件。在白盒水印方法中，验证者需要在版权验证阶段访问可疑模型的内部信息，包括网络结构、网络参数、训练权重等。验证者通过算法解析模型的内部信息来获取相应的水印信息，从而实现版权认证。相比之下，在黑盒水印方法中，验证者无法访问模型的内部信息。这类模型水印方法常见于只能通过网络应用接口（Application Programming Interface,

API）访问模型的场景。一般情况下，验证者可以远程访问模型 API 来获得一些输入输出对，再根据这些输入输出对的特殊表现来验证模型的所有权。

表 1.1 版权验证时不同模型水印方法需满足的条件

模型水印方法	访问模型内部信息	访问模型权限	模型输出
白盒水印	√	√	√
黑盒水印	×	√	√
无盒水印	×	×	√

目前，白盒水印和黑盒水印主要应用于分类模型，而无盒水印算法主要用于生成式网络。在版权认证过程中，无盒水印技术无需将水印嵌入到模型的内部参数中，也无需构建特定的输入输出对来充当水印。在这种方法中，受保护模型的所有输出都会携带水印信息。模型所有者可以通过提取输出中的水印信息来验证模型的版权。由于无需模型的参与，该方法被称为无盒水印。

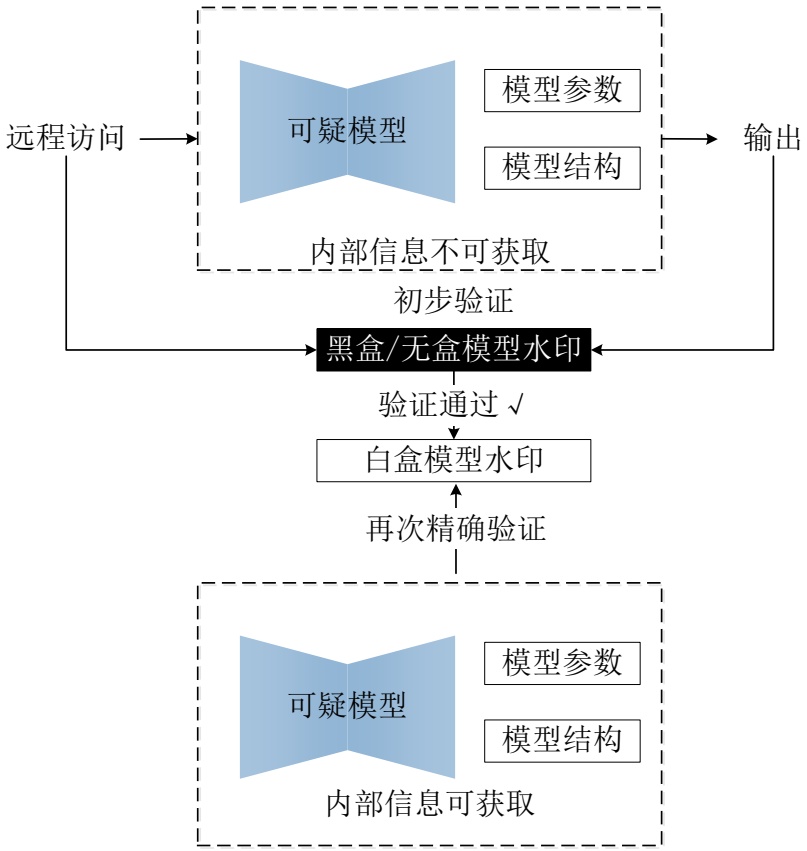


图 1.1 多种模型水印模式配合验证

在版权验证的过程中，上述的模型水印方法也可以相互配合。如图 1.1 所示，验证者可以利用黑盒水印或无盒水印方法，通过访问远程模型的 API 或输出来进行初步的版权认证。一旦初步验证通过，验证者就可以要求执法部门强制公开可疑模型的内部信息。随后，验证者可以进一步实施白盒水印验证，以实现对深度模型版权的精准认证。

1.2.1 白盒水印

白盒水印意味着在验证时，验证者可以访问深度模型的内部信息。具体而言，秘密水印信息将被嵌入到模型的内部，验证者可以从模型内部提取水印以完成版权验证。

Uchida 等人[14]提出了第一种白盒水印方法，旨在通过标记模型来实现知识产权保护。该方法首先生成一个密钥作为水印信息，然后将水印与原始模型参数进行点乘，从而得到水印模型参数。通过在损失函数中引入正则项约束，可以使原始模型参数分布接近于水印模型参数分布，从而实现了水印信息的嵌入，并保证了模型原始任务的性能不会显著下降。在版权验证的关键阶段，验证者首先从含有水印信息的模型参数中精准地提取出隐含的水印。随后，通过与预先设定的原始水印信息进行细致的比对分析，验证者能够准确判断模型是否侵犯了知识产权。Rouhani 等人[15]提出了一种创新的模型水印方法，该方法建议将特定的水印字符串嵌入到深度学习网络各层的概率密度函数中。这种设计巧妙地使得模型对输入数据计算出的概率密度函数输出本身成为携带水印信息的媒介。通过这种方式，嵌入的水印信息不仅与模型内部的参数紧密相关，而且与输入数据的特征也建立了联系。这种动态关联的实现，为创建一种能够随输入变化而变化的动态水印机制提供了可能，进一步增强了水印的隐蔽性和鲁棒性。基于此，该方法成功地解决了 Uchida 等人的方法[14]在执行验证后可能泄露静态水印信息的问题。因此，该方法也避免了水印覆盖攻击，即攻击者通过二次嵌入水印来破坏版权拥有者的原始水印。Feng 等人[16]采用伪随机方法选择用于嵌入水印的权重参数，以

减少水印信息泄露的风险，并采用了基于补偿机制的微调方法以抵消微小的精度下降。由于水印嵌入位置的隐蔽性，该方案同样克服了水印覆盖攻击。

上述方法都导致了嵌入水印后的模型参数权值分布与原始模型权值分布有区别，很容易被基于统计分析的攻击方法检测出水印的痕迹。并且，嵌入水印信息越大，权重的差异越明显，越容易被攻击者发现。为了解决以上白盒水印技术缺乏隐蔽性的问题，Wang 等人[17]结合对抗训练的方法，强制约束含水印模型的参数分布与原始模型的参数分布一致，从而提高了模型水印的隐蔽性和鲁棒性。Cortiñas-Lorenzo 等人[18]发现水印嵌入过程中使用不同的优化算法也可能会导致模型的权值分布发生显著变化。为了解决这个问题，Cortiñas-Lorenzo 等人[18]提出了一种称为块正交投影的新方法，该方法能将水印与 Adam 优化器相结合，从而提升了水印的隐蔽性和鲁棒性。Li 等人[19]则运用了传统数字水印中的量化调制和扩频手段，将水印嵌入到模型权重的扩频伪随机序列的投影中，从而提升了白盒水印的容量和隐蔽性。然而，仅有基于微调的水印嵌入方案适用于此方法。Chen 等人[20]为了解决共谋攻击问题，采用了抗合谋攻击生成码生成水印信息，最后结合现有的白盒水印算法完成水印信息的嵌入。

除了利用深度模型的权重参数作为水印载体之外，深度模型的结构信息也可用作水印载体。同时，基于模型结构的水印方法对参数修改攻击具有固有的鲁棒性。Luo 等人[21]利用神经结构搜索技术来获取所需的水印结构。神经结构搜索技术可以自动寻找最优网络架构，这些架构足够独特，可以代表所有权。在验证阶段，验证者可以通过基于侧通道的高保真模型技术来提取此类水印。然而，Luo 等人[21]的方法只适用于重头训练模型的场景，无法在已经训练完成的模型上实现水印嵌入。为了克服这一局限性，Zhao 等人[22]提出了一种新颖的结构水印技术，在水印嵌入过程中，使用水印控制的通道修剪率来修剪深度模型的内部通道。在水印提取过程中，通过从目标深度模型的架构中识别通道剪枝率来检索水印。由于剪枝机制的优越性，该方法在水印嵌入过程中保留了模型在其原始任务上的良好性能。

水印混淆攻击，即攻击者利用伪造的水印来声称对模型的所有权。为了应对这类问题，Fan 等人[23]提出了基于护照层的新型深度模型所有权验证方案，该方案不仅可以抵抗网络修改，还能够抵抗水印混淆攻击。嵌入数字护照的要点在于设计和训练深度模型，使得深度模型的性能会在伪造护照的影响下迅速恶化。版权验证时，只有使用正确的真护照，深度模型的性能才不会降低。基于此，可以通过模型性能的优劣来判断模型的版权。

经过近些年的发展，白盒水印不再仅仅是概念，而是已经演变成可以用来保护不同领域模型的有效手段。例如，白盒水印已经成功应用于各种识别、分类和标注任务中[24-26]。 通过嵌入不易被察觉的水印，可以在不影响模型性能的前提下，有效追踪和证明模型的所有权。

1.2.2 黑盒水印

白盒水印要求模型所有者在验证会话期间掌握可疑模型的内部信息（例如结构、参数等），从而提取完整的水印并完成版权验证。但是，这样严苛的条件严重限制了白盒水印的应用范围。为了克服这一挑战，许多学者提出了各种黑盒方法，从而验证者不再需要详细了解模型的内部参数就可以完成水印验证。由于深度模型具备完成特定任务的能力，黑盒水印可以利用深度模型的功能来实现水印嵌入和水印验证。在这个方向上，许多方法通过一组精心制作的样本（称为触发样本）来微调深度模型并标记它，然后通过分析模型对一组触发样本的预测结果来检索嵌入的水印。因此，黑盒水印在验证阶段无需访问模型内部信息，而是只用访问模型的输出或调用模型 API 接口即可。

Adi 等人[27]提出了第一个基于后门攻击的黑盒模型水印方法，该方法采用后门技术来确保深度模型知识产权的安全。具体而言，Adi 等人[27]利用了深度模型过度参数化后容易受到后门攻击的弱点，将弱点转化为深度模型版权保护的优势，从而实现了黑盒水印。为此，验证者将随机选择一些无关图像并打上水印标签来生成触发集。随后，验证者将精心构造的触发集巧妙地整合进原始的训练数据集

中，以便用于后续的深度模型训练。通过这种融合策略，所训练出的深度模型能够在接收到常规输入时保持预期的性能表现，同时，一旦接触到那些特定触发图像，模型便能够精准地识别并输出相应的水印标签。在验证阶段时，验证者无需访问模型的内部信息，而是只需远程访问模型并输入触发集，若触发集的输出能够获取对应的水印标签，则表明水印提取成功，从而验证者可以宣称对模型的所有权。

基于上述研究，Zhang 等人[28]进一步探索了构建不同类型触发集的可能性。具体而言，Zhang 等人[28]尝试了三种构建触发集的方法：文本触发器、噪声触发器和不相关样本触发器。然而，这三种方法也存在一定的缺陷：添加噪声触发器虽然不容易被攻击者察觉，但无法很好地证明所有者的身份；添加文本触发器可能无法绕过人工检查或逃避攻击；而选择无关样本的方法则容易被攻击者发现。类似地，Guo 等人[29]采用了后门技术来保护嵌入式系统的知识产权。Guo 等人[29]提出的方法根据用户信息来生成噪声，从而建立了噪声与模型拥有者之间的对应关系，进而明确了模型的所有权。随后，Guo 等人[30]进一步采用了遗传算法，对触发器的最佳模式及其在模型中的最优嵌入位置进行了系统化的搜索和优化。这一优化过程显著提升了水印系统的整体性能，增强了其在实际应用中的稳定性和鲁棒性。

除了使用触发集的方式实现黑盒水印以外，Merrer 等人[31]通过微调模型决策边界来实现水印。具体而言，Merrer 等人[31]的工作首先集中在识别两类关键的对抗样本：一类是真实对抗样本，即那些被原始模型错误分类的实例；另一类是错误对抗样本，指的是尽管被设计为对抗性，但被原始模型正确分类的样本。这两类样本均位于原始模型决策边界的邻近区域。为了在原始模型中嵌入水印，研究者通过对抗性训练的策略，进一步促使真实对抗样本被正确分类，即按照其真实的标签进行分类。通过这种方式，微调后的模型决策边界本身就构成了一种隐蔽且有效的水印，为模型的版权保护提供了新的技术手段。基于 Merrer 等人[31]提出的方法，Chen 等人[32]引入了一种模型相关的编码方案，并将所有者的二进制签名合并到输出中。具体而言，Chen 等人[32]将模型预测类别平均分为两组，

分别代表“0”和“1”。真正的对抗样本被错误分类属于“0”，则可以表示比特信息“0”，同理可得比特信息“1”。然而，关键样本的错误分类将不可避免地影响模型对原始任务的决策边界。

上述的黑盒水印技术的嵌入容量都是 0 位的，即只能表示含/不含水印。为了增强水印的嵌入容量，Sakazawa 等人[33]提出了一种多位水印方法。具体而言，Sakazawa 等人[33]训练了一个维度拓展模型，并将返回的一维预测值转化成二维图像。触发集图像可以得到一组输出图像，最终的版权图像就通过这些触发集的输出累加得到。

总体而言，黑盒水印成功克服了白盒水印在应用场景上的局限，展现出更强大的实用性和更广泛的应用价值。近年来，随着模型水印作为“盾”的不断发展，各式各样的攻击“矛”也是层出不穷。针对这些威胁，研究者们开发了专门针对特定攻击的模型水印技术，如代理模型攻击[34-35]、查询修改攻击[36]以及水印覆盖攻击[37-38]等。此外，除了传统的基于后门触发集的黑盒水印外，对抗样本[31-32]和模型自身的决策边界[39]也被开发为水印载体，用于在黑盒条件下进行版权认证。随着黑盒水印技术的日趋成熟，其应用也不再局限于图像分类模型，而是已经扩展到图像处理网络、图神经网络等多个领域[40-42]。

1.2.3 无盒水印

无盒水印算法主要应用于生成式网络。对于生成式网络而言，其输出为相应的图像、文字等产物，而不是分类模型那样的分类标签。在生成式模型的版权认证过程中，验证者无法使用触发集访问模型并获取对应的水印标签，因此一般黑盒水印场景下的方法不再适用于生成式模型。另外，验证者也无法直接访问模型的内部参数并从中提取水印，因此一般白盒水印场景下的方法也不适用。实际上，可以将无盒水印视为黑盒水印在生成式网络上的特殊情况。无盒水印只需要获取可疑模型的一些输出即可完成版权验证。具体而言，在模型输出后，所有的输出都将携带水印信息。在验证时，验证者直接对模型的输出进行水印提取或统计分

析，从而完成深度模型的版权认证。因此，无盒水印的版权保护不再需要模型本身的参与。截至目前，无盒水印方法主要保护了图像分割、图像着色、超分辨率等图像任务模型[43-46]。

Wu 等人[43]提出了一种无盒水印框架，该框架将水印损失合并到总损失函数中，从而成功地将水印嵌入到被保护模型的输出中。该水印框架的核心架构由两个关键组件构成：一是作为保护对象的主机网络（即原始被保护的网络），二是受密钥控制的水印提取网络。在训练阶段，这两个网络将利用特定的水印损失函数进行联合优化。通过这种优化策略，不仅确保了主机网络在完成训练后能够维持其原始的图像处理功能，同时也赋予了网络一个新的能力：即允许授权的验证者利用密钥从网络的输出图像中准确检测出嵌入的水印信息，从而实现了对图像内容和模型本身双重保护的目标。此外，该框架还具备强大的版权验证机制：任何未经标记的图像，或使用不正确密钥进行的版权验证尝试，都会导致水印提取网络输出无意义的噪声信号，从而使得版权认证过程失败，有效防止了未授权的使用和潜在的侵权行为。该框架还使用了对抗样本来进行对抗性训练，从而提高了水印系统对常见预处理攻击的鲁棒性。

Zhang 等人[44-45]提出了一种新的图像处理网络水印框架，专门针对代理模型攻击。这种攻击需要收集目标模型的输入和输出，以指导代理模型的训练。代理模型攻击的目标是利用输入输出对训练代理模型，从而使代理模型能够实现与目标模型相似的能力。在 Zhang 等人[44-45]提出的方法中，水印嵌入网络将不可感知的水印嵌入到模型的输出中。然后，水印提取网络可以提取带水印的输出以检索版权信息。在代理模型攻击场景中，攻击者只能访问自己的输入和带有水印的输出。因此，攻击者训练的代理网络的输出中也包含版权信息，从而防止了代理模型攻击。然而，需要注意的是，这种方法可能无法抵御某些预处理攻击，因为所需的一致性可能会受到损害。

然而，上述水印方法并未考虑样本图像的不同通道对水印的影响。因此，水印的嵌入性能仍然有限。针对这一问题，Zhang 等人[46]首先分析了嵌入水印信息对不同通道的影响。然后，根据人类视觉系统（Human Visual System, HVS）的

特点，提出了两种基于 HVS 的生成式模型水印方法，从而提高了保真度，并具有良好的通用性。

1.3 本文研究内容和结构安排

1.3.1 研究内容

与白盒水印和黑盒水印不同，生成式模型水印使用图像作为载体来验证所有权。因此，在评估生成式模型水印的保真度时，需要同时考虑图像空间域和频域上的保真度。但是，现有的生成式模型水印算法不可避免地在输出图像的频域中引入高频伪影，从而损害了含水印图像频域的一致性，进而降低了水印系统在频域上的保真度和隐蔽性。实际上，目前的生成式模型水印仅考虑了图像空间域上的保真度，即以不可察觉的方式将水印嵌入到图像空间域中，而往往忽略了嵌入水印对频域上的影响。

高保真的生成式模型水印意味着水印的嵌入对生成网络输出图像的影响极小，图像添加水印前后的差异（图像空间域和频域）几乎无法察觉。高保真度将提高水印系统的隐蔽性，对于水印系统而言，隐蔽性[47]要求嵌入的水印不被任何攻击者感知。如果攻击者无法感知到嵌入的水印，他们可能不会采取任何行动攻击水印，这意味着更好的隐蔽性也可以在一定程度上降低水印遭受攻击的概率，从而提高水印系统的安全性。但是，高频伪影的存在导致了当前生成式模型水印的保真度降低，并进一步损害了隐蔽性和安全性。综上所述，存在迫切的需求来解决高频伪影问题，从而实现高保真的生成式模型水印技术。基于此，本文建议同时在图像空间域和频率域上考虑生成式模型水印的保真度，并专注于消除频域中的高频伪影，以提升生成式模型水印的保真度、隐蔽性和安全性。围绕上述观点，本文将展开下面两个方面的研究：

1) 基于小波变换的生成式模型水印：为了解决水印嵌入过程导致的高频伪影问题，本文提出了一种基于离散小波变换的框架。该框架利用小波变换建立小

波频域分离层，从而使得目标模型输出的图像被分解为不同的频率成分。再通过对水印嵌入网络和水印提取网络进行联合训练及损失函数的优化，能有效地将水印嵌入到图像的低频区域中，这显著减少了高频伪影，从而实现了高保真的生成式模型水印技术。

2) 基于频域扰动的生成式模型水印：尽管上述研究很好地消除了水印嵌入过程所导致的高频伪影，但是它却没有处理嵌入网络结构可能导致的高频伪影。因此，本文进一步提出了一种新的生成式模型水印框架，专注于消除水印嵌入过程和嵌入网络所引起的高频伪影。具体而言，该方法摒弃了传统的水印嵌入网络，以避免由此产生的固有高频伪影。此外，该方法还设计了一个频域扰动生成网络，用于生成低频扰动。这些低频扰动被作为水印添加到载体图像的低频成分中，从而大幅减少了水印嵌入过程对图像高频特性的影响，最终实现了高保真的生成式模型水印技术。

1.3.2 结构安排

本论文旨在消除生成式模型水印中的高频伪影，从而实现高保真的生成式模型水印技术。为了实现上述目标，本论文开展了两项主要研究。首先，本论文认识到生成式模型水印在频域上存在的高频伪影问题，这一问题显著降低了水印系统的保真度。为应对此问题，本文首先提出了基于小波变换的生成式模型水印，以抑制由水印嵌入过程引起的高频伪影。然后，本文又提出了一种新的基于频域扰动的生成式模型水印算法，该算法有效地消除了由水印嵌入过程和嵌入网络所引起的高频伪影，从而实现了高保真的水印算法。本学位论文共分为五章，各章内容及其逻辑关系如图 1.2 所示。

第一章 绪论：本章主要介绍了模型水印的研究背景、意义及国内外研究现状。此外，本章还总结并分析了本文的创新点和结构安排。

第二章 相关技术基础：本章为后续章节提供必要的背景知识和基础算法。章节内容包括介绍生成式模型水印的定义、生成式模型水印的评价指标、生成式

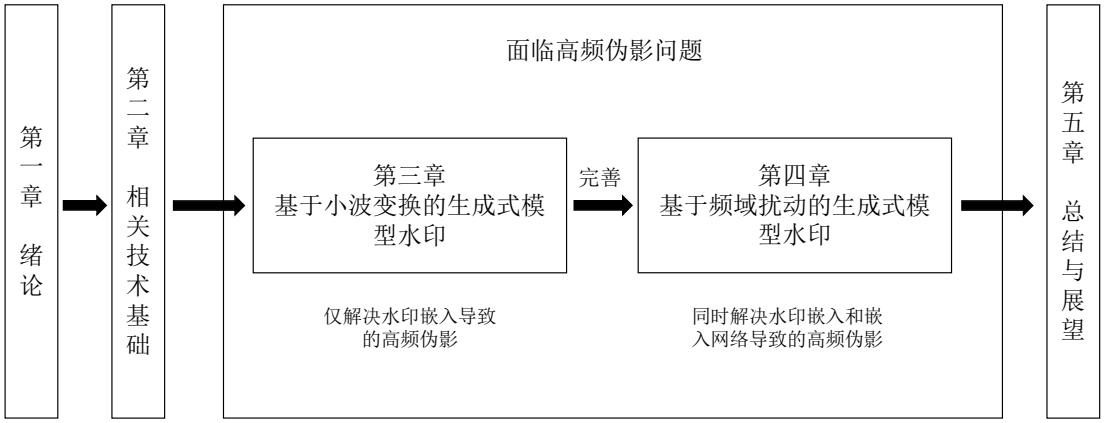


图 1.2 本文各章内容及其逻辑关系

模型水印面临的问题、离散余弦变换和高频伪影的成因分析。以上内容为后续研究提供了理论支持。

第三章 基于小波变换的生成式模型水印：本章首先介绍了当前生成式模型水印中存在的高频伪影问题。为了解决由水印嵌入过程引发的高频伪影，本章提出了一种基于离散小波变换的生成式模型水印框架。此框架能有效地缓解高频伪影问题，并提高生成式模型水印的保真度。框架的核心设计包括使用小波频域分离层将嵌入网络输出的图像划分为不同的频率成分，随后通过联合训练及损失函数优化的过程，水印被有效嵌入到输出图像的低频区域中。综合实验结果显示，该框架成功的抑制了高频伪影。

第四章 基于频域扰动的生成式模型水印：虽然第三章的方法成功抑制了由水印嵌入过程引起的高频伪影，但是它未能解决由嵌入网络引起的高频伪影问题。为应对这一挑战，本章提出了一种生成式模型水印算法，该算法不依赖传统的嵌入网络，从而消除了嵌入网络所引入的高频伪影。该方法使用频域扰动生成网络来生成扰动，这些扰动将作为水印被嵌入到载体图像的低频成分中。实验结果表明，本章的方法有效地消除了高频伪影，从而实现了高保真的生成式模型水印。

第五章 总结与展望：本章主要总结了本文的研究内容，并对该领域的未来发展进行了展望。

1.4 本章小结

本章综合分析了模型水印技术的研究背景与意义，并详尽回顾了国内外研究人员在模型水印领域的进展。本章特别强调了当前生成式模型水印技术中普遍存在的高频伪影问题，这一问题显著影响了水印系统的保真度。此外，本章详述了本文的研究内容与结构安排，旨在帮助读者充分理解本文所探讨问题的重要性及研究的核心内容。这一综合性概述不仅为读者提供了对模型水印技术的全面认识，也为后续章节的深入讨论奠定了基础。

第二章 相关技术基础

2.1 生成式模型水印

2.1.1 生成式模型水印简介

不同于分类模型将输入映射到分类标签，对于生成式网络而言，其输出为相应的图像、文字等产物。实际上，在生成式模型水印的版权认证过程中，验证者无法通过接近模型内部信息或利用触发集访问模型来提取对应的水印。因此，一般的模型水印方法不再适用于生成式模型。为此，研究人员开发了旨在保护生成式模型的水印技术，即生成式模型水印。实际上，生成式模型水印技术可被视为无盒水印技术保护图像生成式网络时的特殊情况。

对于生成式模型水印而言，只需要获取可疑模型的一些输出即可完成版权的认证。具体而言，生成式模型水印所保护的主机网络的所有输出都将携带水印，版权验证时可以直接对主机网络的输出进行提取或分析来得到水印信息，从而完成深度模型的版权认证。生成式模型水印不再需要像白盒水印一样将水印信息嵌入模型内部参数，也不需要像黑盒水印一样构建一些特殊的输入输出对来表示水印。换句话说，它不再需要模型本身的参与。在模型输出后，模型所有的输出都将携带水印信息，通过提取输出中的水印信息，可以验证模型的版权。

本研究专注于保护与图像任务相关的生成模型，如图像分割、图像去雨、图像着色等。设定一个执行图像任务的主机网络 H ，其输入为 $\{x_1,x_2,\ldots,x_n\}$ （属于图像域 X ），输出为 $\{y_1,y_2,\ldots,y_n\}$（构成图像域 Y ）。该网络通过有监督学习完成 $X\to Y$ 的图像任务。引入生成式模型水印算法 W ，该算法旨在给所有图像域 Y 的图像加水印，即 $Y'=W(Y)$（构成图像域 Y' ），同时优化图像域 Y 与 Y' 之间的差异 D，使得 $D(Y,Y')\to 0$。这样，外界用户在输入 X 域的图像后，将只

能通过远程 API 访问到含水印的Y'域图像，也就是说，主机网络分发出来的图像实际上都带有水印。假设发生了侵权行为，模型拥有者可以通过提取Y'域图像中的水印来验证这些图像和主机网络H的版权。图 2.1 展示了生成式模型水印框架的示意图。

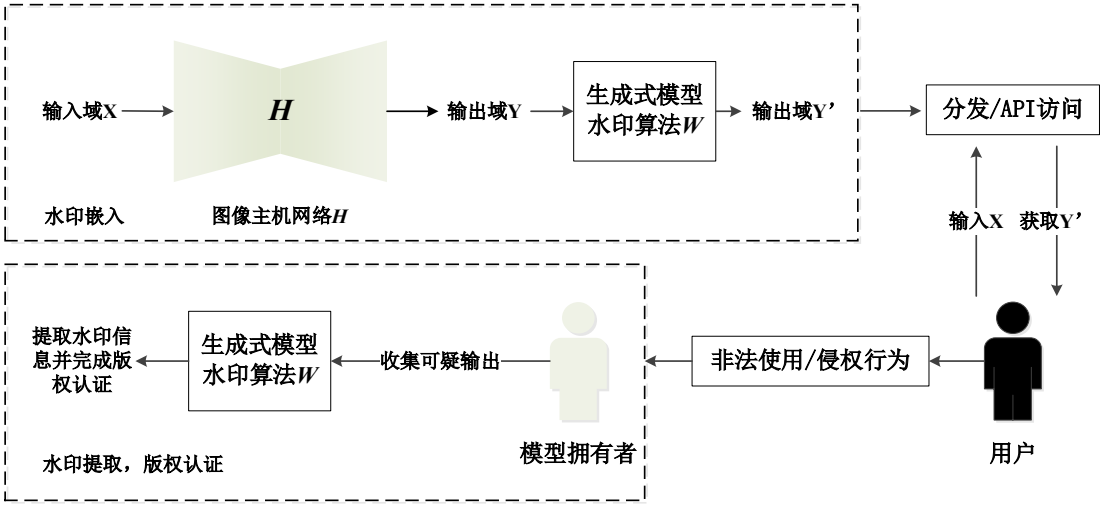


图 2.1 生成式模型水印框架示意图

2.1.2 生成式模型水印的评价指标

在开发生成式模型水印算法时，主要有以下评价指标：

1) 保真度：训练良好的深度模型能够完成特定的任务。模型的性能可能依赖于其特殊的结构或参数，如果随意更改模型以实现水印，会导致模型原始任务无法完成或性能下降，从而使模型失去了保护价值。因此，受保护模型的原始任务性能不应随着水印的嵌入而显著下降。对于生成式模型水印而言，高保真意味着嵌入水印的图像应保持较高的质量，图像嵌入水印前后的差异应极小，无论是在图像空域还是频域上。

2) 隐蔽性：高保真意味着图像嵌入水印前后差异极小，这有效隐藏了水印的存在，进而提升了系统的隐蔽性。对于一个水印系统，隐蔽性要求嵌入的水印不应被任何攻击者感知[47]。如果攻击者无法感知嵌入的水印，则可能不会采取任何行动来攻击水印。这意味着更好的隐蔽性可以在一定程度上降低水印被攻击的

概率，从而提高水印系统的安全性。反之，如果攻击者察觉到了水印，则可能采取各种手段去除或混淆水印，从而导致版权保护失败。

3) 鲁棒性：水印一旦嵌入深度模型中，可能会面临攻击者的恶意破坏及攻击，例如微调、代理模型攻击、预处理攻击等。因此，水印系统需要对特定攻击具备一定的抵抗力，以确保水印系统的有效性。鲁棒性体现了水印系统对抗攻击的能力。对于生成式模型水印而言，含水印的图像是被攻击的目标。一个具备鲁棒性的水印系统应确保，即使含水印图像受到一定的攻击后，仍然能够从中提取出水印，以完成版权认证。

4) 嵌入量：水印作为隐藏的信息被嵌入深度模型中，不同的嵌入载体对水印的嵌入容量有不同的影响。一般来说，较高的嵌入量能嵌入更多信息，但可能导致水印易于被检测，因此需要在嵌入量和隐蔽性之间找到平衡。模型水印追求更高的嵌入量以增强版权保护的效果。例如，零比特水印仅能用于检测水印的存在，而多比特水印能够标识版权所有者。对于生成式模型水印而言，图像作为水印载体具有较高的冗余性，所以能够实现良好的嵌入性能。

5) 可靠性：水印算法应具备高度的置信度以验证版权，即含水印的图像能够以高成功率提取水印，而不含水印的图像则不能提取到任何水印。

6) 安全性：水印系统可能会被攻击者发现和攻击。在实际应用中，水印系统可能会因攻击者的多种攻击手段而变得不再有效。因此，确保水印系统的安全性，保证其版权保护功能的有效性至关重要。安全性的实现依赖于上述指标的良好表现。

7) 计算开销：水印算法同样涉及到深度模型的训练，这个过程也需要消耗计算资源和人力资源等。因此，在实际情况中，要考虑水印算法的开销和受保护模型的价值。

生成式模型水印的高保真可以提高水印系统的隐蔽性和安全性。因此，本文对高保真的生成式模型水印技术进行了研究。具体而言，本文着重研究了高频伪影问题，这一问题导致了水印系统的保真度下降。针对高频伪影问题，本文提出

了两种消除高频伪影的方案，从而实现了高保真的生成式模型水印技术，并提升了水印系统的隐蔽性和安全性。

2.1.3 生成式模型水印所面临的问题

目前的生成式模型水印技术[43-46]主要关注以不可察觉的方式在空间域中嵌入水印，却往往忽视了水印在频域中也需要不可察觉。具体而言，由于大多数生成式模型水印算法都是基于水印嵌入网络来实施水印嵌入过程，因此它们在所保护模型的输出中不可避免地引入了高频伪影。这些高频伪影的存在使得水印在频域上变得可见，从而大幅度降低了水印算法在频域上的保真度。这种频域上的重大缺陷不仅降低了水印算法的整体保真度，还进一步削弱了水印系统的隐蔽性和安全性。因此，现有的生成式模型水印算法正面临着严重的高频伪影问题。

本节进一步明确了空间域伪影和高频伪影的定义。空间域伪影在本研究中被定义为图像中的异常图案，如绿点、红线等。如图 2.2 所示，第一行展示了文献[61]生成的图像（不含水印图像），而第二行则展示了利用文献[43]保护的图像（含水印图像）。可以发现，含水印的图像与不含水印的图像极为相似，这说明

图 2.2 含水印样本的示例

文献[43]在嵌入水印的同时能够保持原始任务的性能，即实现了良好的保真度。然而，也可以从含水印图像中观察到一些轻微的空间域伪影，这些伪影不仅降低

了图像的质量，还可能暴露水印的存在。综上所述，空间域伪影的存在揭示了当前生成式模型水印技术中的一些缺陷。因此，为了生成高质量的图像，应尽量避免这些伪影的出现。

高频伪影通常表现为频域中的不自然、不一致的纹理，如网格状图案。在自然图像中，低频信息通常对应于图像中强度平滑变换的区域，包含了图像的主要信息。而图像中的边缘、像素强度的突变则通常由高频函数描述。因此，自然图像（不含水印）的大部分能量通常集中在低频部分[53-54]。自然图像的 DCT 频域检测结果如图 2.3 所示，可以发现自然图像的 DCT 频域检测结果符合上述观点。另外，为了直观地阐述高频伪影现象，图 2.4 展示了含水印图像的 DCT 频域检测结果。可以发现，图 2.4 的 DCT 频域检测结果中有许多小格状的亮点，这些都被视为频域上异常的图案，即高频伪影。

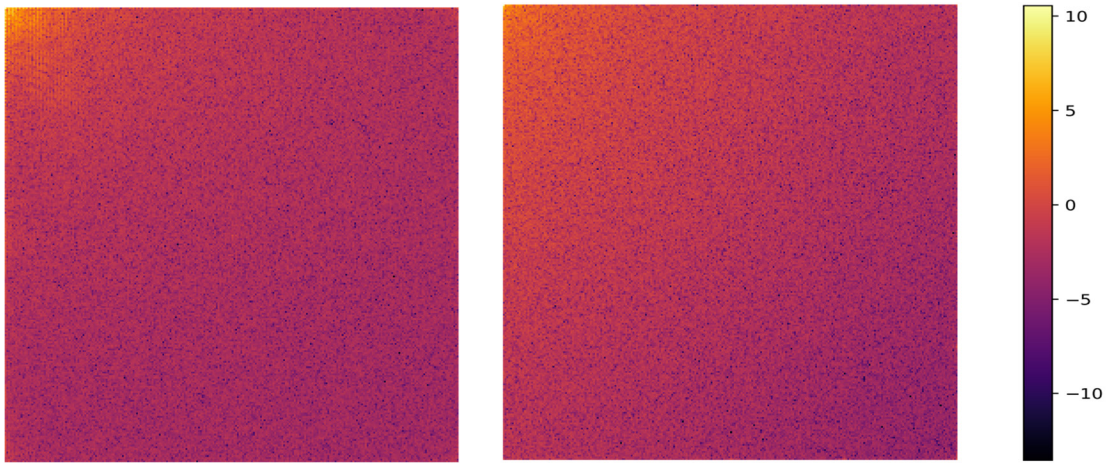


图 2.3 自然图像的 DCT 检测结果

综上所述，本节分析了生成式模型水印技术所面临的主要挑战，即高频伪影问题带来的严重威胁。尽管现有方法能够在空间域中以不可见的方式嵌入水印，大多数情况下不会在空间域产生伪影，但它们普遍忽视了嵌入水印对频域的影响。这种疏忽导致生成式模型水印的频域上出现高频伪影，这些高频伪影不仅降低了图像的质量，更暴露了水印的存在，增加了水印被攻击的风险。因此，解决生成式模型水印中存在的高频伪影问题显得尤为迫切。为此，本文提议在空间域和频

率域上双重优化生成式模型水印的保真度，重点研究如何消除频域中的高频伪影，以实现高保真的水印技术。

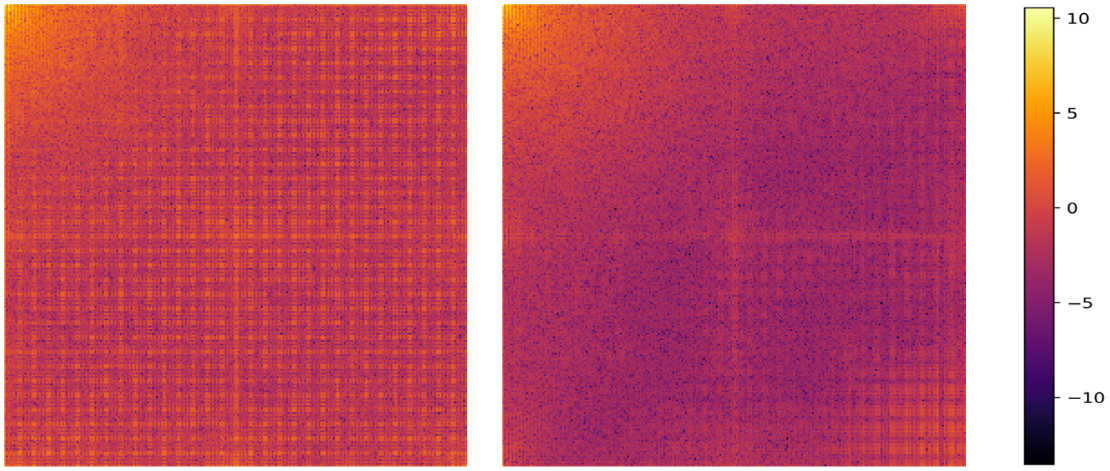


图 2.4 含水印图像的 DCT 检测结果

2.2 离散余弦变换

为了检测高频伪影，本文利用了离散余弦变换（Discrete Cosine Transform，DCT）。DCT 是一种广泛使用的数字信号处理方法，用于将信号从空间域转换到频率域。这种变换将有限序列描述为频率不同的余弦函数的求和。假设有一个矩阵 $A \in R^{n\times n}$，其中每个元素 $a_{i,j}$ 对应一个位置 (i,j)，本文可以应用 DCT 变换得到变换后的矩阵 $B \in R^{n\times n}$。位于位置 (i,j) 的元素 $b_{i,j}$ 可以表示为公式(2.1)：

$$b_{i,j} = c_i c_j \sum_{p=0}^{n-1}\sum_{q=0}^{n-1} a_{p,q} \cos\left(\frac{(2p+1)i\pi}{2n}\right)\cos\left(\frac{(2q+1)j\pi}{2n}\right) \tag{2.1}$$

其中 $i=j=0$ 时，$c_i = c_j = 1/\sqrt{4n}$ 否则，$c_i = c_j = 1/\sqrt{2n}$。A 可由 B 通过公式(2.2)重建，即:

$$a_{i,j} = \sum_{p=0}^{n-1}\sum_{q=0}^{n-1} c_p c_q b_{p,q} \cos\left(\frac{(2i+1)p\pi}{2n}\right)\cos\left(\frac{(2j+1)q\pi}{2n}\right) \tag{2.2}$$

生成式对抗网络（Generative Adversarial Network, GAN）生成的图像在频率空间中可能会表现出明显的高频伪影，这些高频伪影可以通过 DCT 热图进行有效检测[49]。目前的生成式模型水印技术中具有类似于 GAN 的特性，即水印嵌入网络中广泛使用了采样操作，这增加了引入高频伪影的风险。鉴于此，本文检测了生成式模型水印所标记的图像中是否存在高频伪影。具体而言，本文采用了 II 型 2D-DCT[52]将图像从空间域转换到频域，并通过 DCT 热图来检测高频伪影。DCT 热图中，每个点的亮度表示对应频率系数的大小，频率从图像的左上角（低频区）到右下角（高频区）逐渐增高。最终，本研究通过 DCT 热图检测到了生成式模型水印中的高频伪影。

2.3 高频伪影的成因分析

本文重点研究了如何消除生成式模型水印中的高频伪影。为此，本节首先对高频伪影的成因进行了分析。Odena 等人[48]将空间图像中的网格状伪影与上采样联系起来。另外，Zhang 等人[49]还指出，在生成网络中使用的上采样操作会形成从低维潜在空间到高维数据空间的映射，这一映射过程不可避免地在频域中产生网格状高频伪影。基于上述观点，本文认为在生成式模型水印中，水印嵌入网络中频繁的采样操作是导致标记图像的频域中出现高频伪影的主要原因。

此外，Zeng 等人[50]还提到，在后门攻击中使用触发器对图像进行修补时，会直接引入高频伪影，因为触发器本身可能携带固有的高频伪影。此外，由于添加触发器会导致相邻像素之间的相关性降低，也需要一个高频函数来近似修补后的数据。文献[51, 59]的作者发现，当使用深度神经网络（Deep Neural Network, DNN）嵌入信息时，信息将以高频编码的形式添加到载体图像中。因此，本文认为生成式模型水印中嵌入水印的过程与后门攻击或基于 DNN 的信息嵌入过程相似。在生成式模型水印技术中，通过水印嵌入网络将秘密水印嵌入图像时，实际上是在标记图像中添加了一种特殊的扰动。由于时频特性，空域上添加扰动也会影响到频域，从而产生了高频伪影。

综上所述，生成式模型水印中高频伪影出现的原因主要来自于两个方面。一个是水印嵌入网络的固有结构，另一个是水印嵌入过程中所添加的扰动。水印嵌入网络通常涉及大量的上采样和下采样操作，这将会无意中将高频伪影引入输出图像的频域中。同时，在输出图像中嵌入水印的过程类似于引入高频扰动，这也会直接导致图像频域中出现高频伪影。虽然水印嵌入网络和水印添加过程对于生成式模型水印算法的实现至关重要，但这也不可避免的引入了高频伪影。

然而，目前的生成式模型水印算法都通过水印嵌入网络来嵌入水印，再通过水印提取网络来提取水印以实现所有权验证。例如，文献[43, 46]直接利用主机网络作为水印嵌入网络来嵌入水印，再采用提取网络来提取水印；文献[44-45]则引入了额外的水印嵌入网络来嵌入水印，再采用提取网络来提取水印。由于现有的生成式模型水印算法的实现都依赖于水印嵌入网络和水印嵌入过程，所以它们都不可避免的引入了高频伪影。

高频伪影的存在显著降低了生成式模型水印的保真度，使得水印在频域中容易被识别，从而允许攻击者能轻松地在频域中检测并移除水印。因此，迫切需要方案来解决生成式模型水印所面临的高频伪影问题。为此，本文提出了两种高保真的生成式模型水印框架，旨在消除高频伪影的影响，并提高水印系统的保真度和隐蔽性。

2.4 本章小结

综上所述，本章介绍了生成模型水印的相关技术基础。具体而言，本章首先介绍了生成式模型水印技术和生成式模型水印技术的评价指标。然后，本章特别强调了当前生成式模型水印技术所面临的高频伪影问题，该问题显著影响了水印系统的保真度和隐蔽性。随后，本章介绍了离散余弦变换在检测高频伪影中的应用。最后，本章深入探讨了高频伪影的产生原因，并将高频伪影的产生原因归纳为两点。这些讨论为后续研究消除高频伪影并实现高保真的生成式模型水印提供了理论和研究基础。

第三章 基于小波变换的生成式模型水印

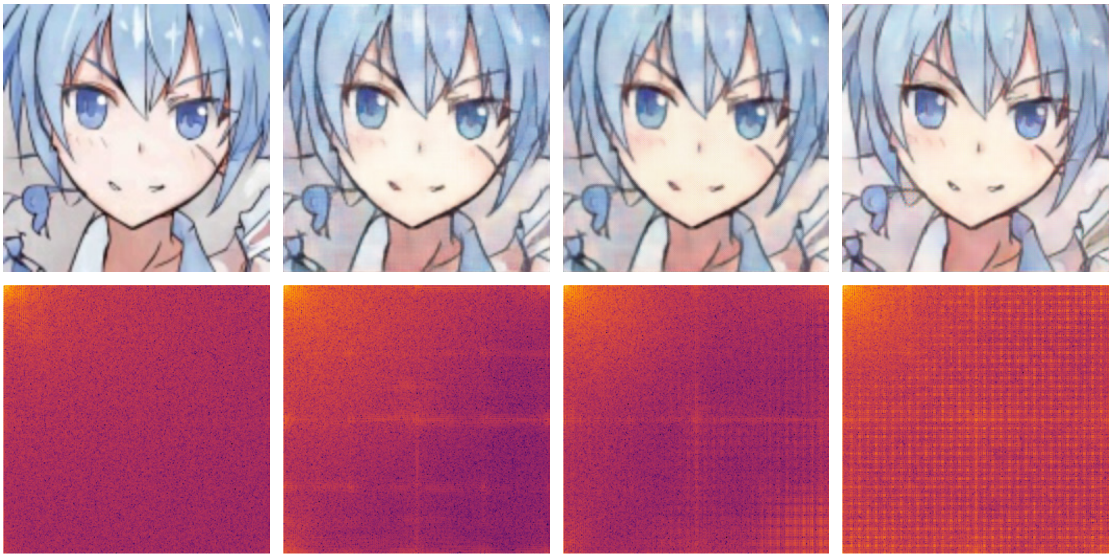
3.1 研究动机

在过去的几十年中，人工智能技术[1-6]与物联网[57, 58]的快速发展极大地促进了人类生活和生产方式的变革。鉴于生成式模型成为了物联网中的重要数字资产，通过生成式模型水印技术保护这些模型的知识产权就显得尤为重要。现有的生成式模型水印技术通过嵌入水印到主机网络的输出中来实现版权保护，而嵌入过程的实现通常依赖于深度神经网络。本文 2.1.3 节中分析了当前生成式模型水印技术所面临的高频伪影问题，这值得引起研究人员的重视。高频伪影的存在使得水印在频域上容易被察觉，大大降低了水印算法在频域上的保真度。尽管当前的生成式模型水印算法在图像空间域上实现了较高的保真度，但在频域上保真度很低，这使得攻击者能够轻易地发现水印存在，从而有可能消除水印。因此，现有生成式模型水印中存在的高频伪影问题迫切的需要解决方案。

为此，本章建议同时从空间域和频率域考虑生成式模型水印的保真度，并专注于抑制生成式模型水印在频域上的高频伪影，以提高生成式模型水印的保真度。换句话说，本章的目标是通过抑制生成式模型水印中的高频伪影，以实现高保真的生成式模型水印技术。

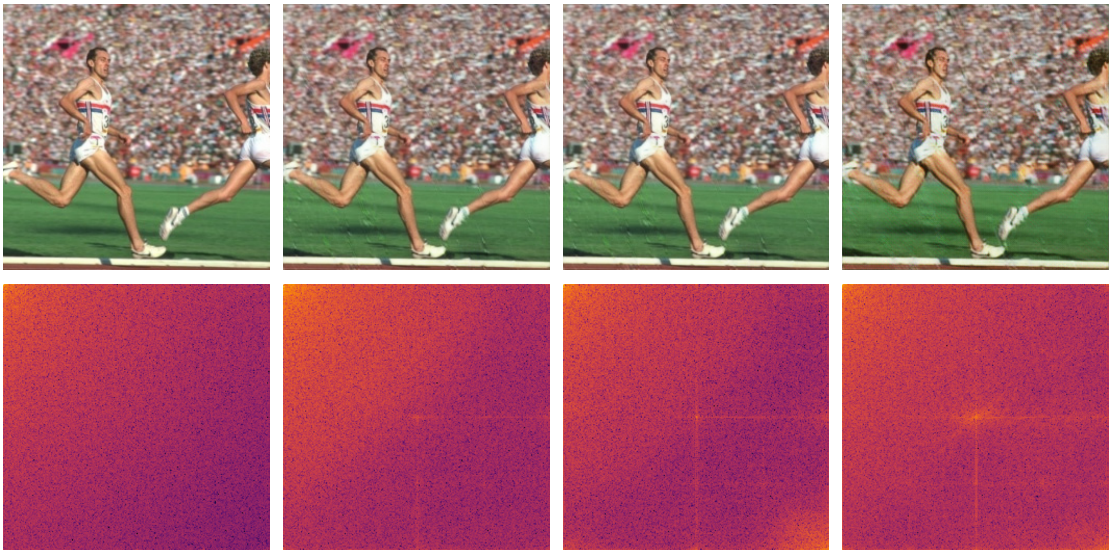
为了进一步展示生成式模型水印所面临的高频伪影问题，本节在 de-raining 数据集[61]和 Danbooru2019 数据集[60]上进行了实验。具体而言，本节在上述数据集上检测了文献[43-44]中的高频伪影。图 3.1 和图 3.2 分别显示了在这两个数据集上的高频伪影检测结果。图 3.1(a)是随机从 Danbooru2019 数据集中提取的地面实况图像，图 3.1(b)是由文献[56]所生成的图像。图 3.1(c)(d)分别是文献[43-44]标记的图像。可以发现，图 3.1 中的四幅图像在视觉效果上相互接近，这说明文献[43-44]都能在不影响文献[56]的原始任务的情况下完成水印任务。不过，通过仔细观察，还是可以发现生成的标记图像有轻微的空间伪影。例如，在图 3.1(d)

中，可以看到空间域的一些异常状态（如绿点、红线等）。因此，为了提高水印系统在空间域上的保真度，应该避免这些异常状态。图 3.2 显示了类似的结果。



(a) 地面实况　(b) 不含水印图像[55]　(c) 含水印图像[43]　(d) 含水印图像[42]

图 3.1 paint transfer 任务的示例



(a) 地面实况　(b) 不含水印图像[55]　(c) 含水印图像[43]　(d) 含水印图像[42]

图 3.2 de-raining 任务的示例

更重要的是，本节进一步检测了测试图像在频域上的特征。具体而言，本节利用了 2.2 节中提到的 DCT 热图来检测测试图像的高频伪影。图 3.1 和图 3.2 的第二行分别显示了不同方法相应的 DCT 频域检测结果。一般来说，对于自然图像，低频信息是图像强度平滑变换的地方，代表了图像的大部分信息。相反，图

像中的边缘代表像素的突然变化，应该用较高频率的函数来近似。因此，自然图像的大部分能量通常都集中在低频分量上[53-54]。图 3.1(a)所给出的结果与这一结论一致，即只有左上角明亮，其他地方偏暗，这说明能量集中于低频区域（DCT 热图上越明亮代表能量越大）。本文第 2.1.3 节中定义高频伪影是指频域中一些不自然和不一致的纹理（如格子状图案）。如图 3.1(b)所示，可以清楚地看到格子状的高频伪影。如图 3.1(c)(d)所示，可以发现更严重的格状高频伪影。图 3.2 展示了类似的结果。因此，通过 DCT 频域检测，本节揭示了生成式模型水印中严重的高频伪影问题。

3.2 方案设计

3.2.1 理论分析

为了降低水印嵌入过程对标记图像高频区域的影响，从而提高水印系统的保真度，一种直观的做法是尽可能将水印嵌入图像的低频区域。为实现这一目标，本章提出了一种基于小波变换的生成式模型水印框架。具体而言，本章构建了一个基于小波变换的频率分离层，用于将嵌入网络生成的图像分解为不同的频率成分；再通过优化联合损失函数来将水印嵌入到图像的低频区域；最后联合训练嵌入网络和提取网络，以确保水印提取网络能从图像低频区域中提取水印。根据时频一致性原理，将水印嵌入低频区域可以显著减少对图像高频区域的影响，从而达到消除高频伪影的目的。

为了进一步证明限制水印嵌入位置在减少高频伪影方面的有效性，本节设计了一个实验，其示意如图 3.3 所示。具体而言，本节首先训练了一个生成模型[56]，然后利用频域分离层对生成式模型所产生图像的频率成分进行分解，以此将水印分别嵌入图像的低频、中频和高频。最后，通过对嵌入网络和提取网络进行联合训练，以确保水印提取网络从相应的频率成分中提取出水印。实验结果如图 3.4 所示。图 3.4(a)显示了无限制情况下直接在全域嵌入水印后的 DCT 检测结果。

图 3.4(b)(c)(d)分别显示了只在低频域、中频域和高频域嵌入水印后相应的 DCT 检测结果。值得注意的是，图 3.4 (c)中部（即中频区域）出现了一些小块状的高频伪影；图 3.4(d)的右下角（即高频区域）充满了小块状的高频伪影；图 3.4(b)中没有小块的高频伪影。这些结果表明，将水印限制在图像的特定子带确实会影响对应的频域。另一方面，如果不限制嵌入位置，水印很容易影响整个频谱，如图 3.4(a)所示，整个频域都充满了块状伪影。总的来说，将水印限制在低频子带可以更好地抑制高频伪影，这表明了限制水印嵌入位置对提高水印系统保真度的有效性。

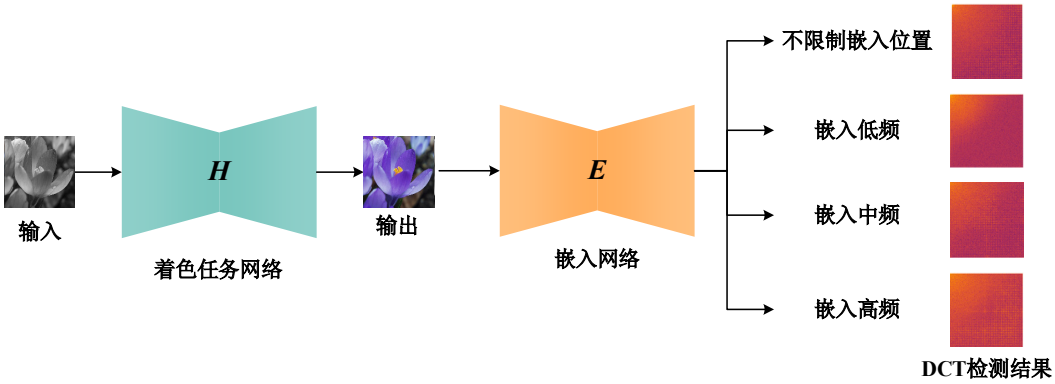
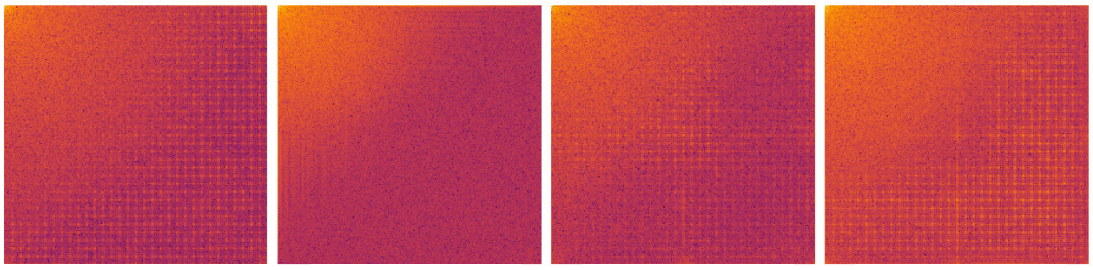


图 3.3 简易实验示意图



(a)不限制嵌入位置　(b)嵌入低频　(c)嵌入中频　(d)嵌入高频

图 3.4 DCT 检测结果

3.2.2 框架概述

目前的生成式模型水印技术可以分为两类。一类方法是直接将主机网络作为水印嵌入网络[43, 46]，主机网络需要同时完成图像任务和水印任务。另一类方法是使用独立的水印嵌入网络来完成水印任务[44-45]，即，主机网络完成图像任务，水

印嵌入网络完成水印任务。与前一类方法相比，后一类方法加速了模型的训练过程，使得主机模型的收敛速度更快。为了利用这一优势，本文采取了第二种策略。

如 2.3 节所分析，水印嵌入过程导致了当前的生成式模型水印技术在频域中引入了高频伪影，这不可避免地降低了水印系统的保真度。因此，为了提高生成式模型水印的保真度，本章提出了一种基于小波变换的生成式模型水印框架。本章提出的生成式模型水印框架如图 3.5 所示。拟议的框架由五个主要模块组成，即：主机网络 H，水印嵌入网络 E，小波频率分离层 F，水印提取网络 R 和鉴别器 D。H 的任务是执行与图像处理或图像生成相关的任务，如 paint transfer[56]，风格转移[55]，语义分割[56]，de-raining[61]。一般来说，H 接受一个图像（或必要时多个图像）作为输入，并生成一个图像作为输出。

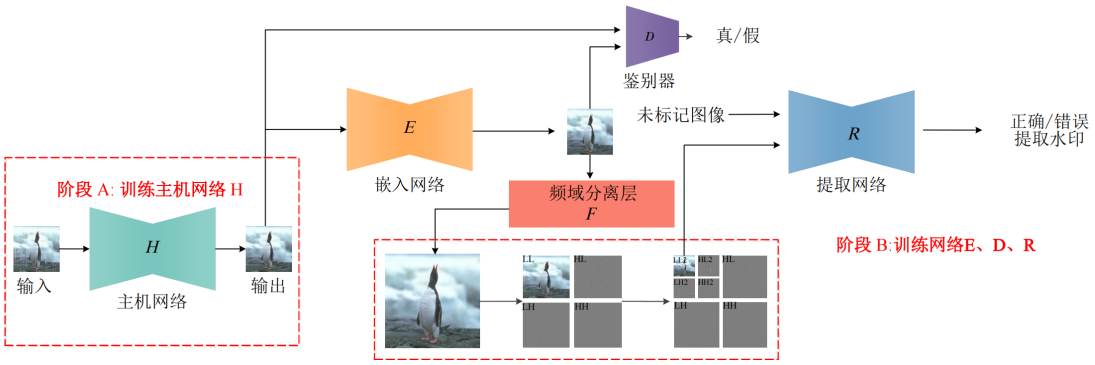


图 3.5 基于小波变换的生成式模型水印框架图

在训练阶段 A，主机网络 H 将完成原始图像任务。在训练阶段 B，嵌入网络 E、鉴别器 D 和提取网络 R 一起被训练。具体而言，水印嵌入网络 E 将接收主机网络 H 生成的图像并对其进行标记。频率分离层 F 负责提取标记图像的不同频率成分，并将标记图像的低频部分 LL_2 发送给水印提取网络 F。注意，如果频率分离层 F 位于嵌入网络 E 之前，生成的标记图像将会丢失较多的中高频信息，从而导致生成的标记图像质量较差。然后，水印提取网络 R 将从标记图像的低频成分中提取水印。为了防止 R 过度拟合，本章将含/不含水印的图像都放入 R 中进行训练，只有含水印的图像才能提取出正确的水印，而不含水印的图像则会提取出噪声。在训练过程中，嵌入网络和提取网络是联合训练的。这种训练方法确保

了嵌入网络能将水印嵌入到生成图像的低频部分，同时使得水印提取网络能够从相应的低频部分提取出水印。

值得注意的是，在进行水印提取之前，含水印的图像可能会遭受各种图像预处理攻击。为了确保水印系统能够抵御常见的预处理攻击，最流行的解决方案是在模型训练期间引入额外的对抗训练阶段，以模拟真实的攻击情况[43]。因此，本章在嵌入网络和提取网络之间增加了一个攻击模拟层，用于进行不同的预处理攻击对抗训练。这样，即使含水印的图像受到攻击，提取网络仍然能够成功从中提取出水印。

3.2.3 结构设计

如第 3.2.2 节所述，本章所提出的生成式模型水印算法由五个模块组成，在这一节中，将对这五个模块的详细结构进行说明。主机网络 H 的架构设计没有严格的限制，只要能高效地执行图像处理任务即可。对于鉴别器 D，本章选择使用 PatchGAN[56]，这是一种被广泛认可的网络架构，以其卓越的性能而闻名。对于水印嵌入网络 E，本章采用了论文[55]中类似 U-Net[6]的网络架构，并将输入和输出维度修改为 $256\times256\times3$，以匹配所需的输入和输出维度。对于水印提取网络 R，本章采用了 Cycle-GAN[62]中类似 ResNet 的网络架构。

此外，为了分离图像的不同频率成分，本章利用了离散小波变换（Discrete Wavelet Transform, DWT）[63]来建立频率分离层 F。DWT 是一种功能强大的数学工具，可将信号或图像分解为不同尺度和频率的子信号或图像，从而捕获信号或图像在不同尺度上的详细信息。这种方法既能保留信号或图像的全局特征，又能捕捉局部特征和边缘信息。通过 DWT 分解得到的子图像被称为子带，通常包括四个子带，即 LL、HL、LH、HH。值得注意的是，LL 子带还可以通过 DWT 进一步分解，得到多个不同频率的子带（即 LL_2、HL_2、LH_2、HH_2）。为了提取所需的低频成分，本章对图像进行了两次 DWT 变换，最终将水印嵌入了 LL_2 子带。

3.2.4 损失函数

拟议的框架包括两个训练阶段。阶段 A 是训练主机网络 H 来完成原始任务。而阶段 B 则是训练其他网络完成水印任务。这两个阶段是相互独立的，因为同时从零开始训练所有网络不能保证模型训练的快速收敛。然而，必须承认，如果能找到一种有效的联合训练策略，可以将这些网络一起训练。

阶段 *A*：主机网络 H 负责完成图像处理任务或图像生成任务。虽然损失函数取决于特定的任务，并且可以有很大的变化，但它通常旨在最小化生成的图像与真实图像之间的距离。为了提高生成图像的视觉质量，还可以使用对抗损失函数。但为简单起见，本章使用文献[56]中使用的深度模型作为主机网络 H，paint transfer 任务使用与文献[56]相同的训练策略，de-raining 任务使用与文献[61]相同的训练策略。训练完成的主机网络 H 就成为了模型水印技术所要保护的对象。

阶段 *B*：为了实现模型水印任务，水印嵌入网络 E、鉴别器 D、水印提取网络 R 三个神经网络将从头开始训练。首先，本章希望提取网络 R 可以从嵌入网络 E 所标记图像的低频成分 LL_2 中提取出水印，并且从不含水印的图像中提取出随机噪声。为此，本章设计了水印提取的损失函数。具体而言，通过阶段 A 的训练后，主机网络 H 可以有效的完成图像处理或图像生成任务。在训练阶段 B，嵌入网络 E 应该可靠地将水印嵌入到主机网络 H 所生成图像的 LL_2 低频分量中。并且，提取网络 R 应该能从水印图像的低频成分中提取出水印。为了实现上述提取水印的目标，可以最小化公式(3.1)：

$$\mathcal{L}_1 = \frac{1}{|N|}\sum_{x\in S_{\text{in}}}\left\| R\left(F\left(E\left(H\left(x\right)\right)\right)\right) - w\right\|_1 \tag{3.1}$$

其中，N 为图像总像素数，$x \in S_{\text{in}}$ 表示主机网络 H 的输入域。$H\left(\cdot\right)$ 表示主机网络 H 所生成的图像，$E\left(\cdot\right)$ 表示嵌入网络 E 所标记的图像。$F\left(\cdot\right)$ 表示频率分离层所提取的 LL_2 低频分量。$R\left(\cdot\right)$ 表示嵌入网络 R 所提取的水印，w 表示地面真实水印。

最终，通过优化公式(3.1)，提取网络 R 能够成功的从嵌入网络 E 所标记的图像的低频分量中提取出水印。

另一方面，为了保证提取网络 R 的可靠性，防止过拟合。提取网络 R 应该从不含水印的图像中提取出噪声，这启发了本章最小化公式(3.2):

$$\mathcal{L}_2 = \frac{1}{2|N|}\sum_{x\in S_{\text{in}}} \left\| R(x) - w_z \right\|_1 + \left\| R\left(H(x)\right) - w_z \right\|_1 \tag{3.2}$$

其中 w_z 是随机噪声。通过优化上述公式(3.2)，提取网络 R 将从不含水印的图像中提取出随机噪声。除非另有说明，本章将默认使用 L_1 范数作为距离度量。最终，水印提取的损失函数可以表示为: $\mathcal{L}_{ext} = \mathcal{L}_1 + \alpha\mathcal{L}_2$。其中 α 是可调超参数。

除了完成水印任务，本章的研究重点是保证水印系统在图像空间域和频域上的高保真。因此，需要保证在图像空间域上以不可见的方式嵌入水印，使得图像添加水印前后在空间域上的变化最小。为了实现该目的，本章希望最小化主机网络 H 生成的图像与嵌入网络 E 所标记的图像之间的差距，可以表示为公式(3.3)：

$$\mathcal{L}_3 = \frac{1}{|N|}\sum_{x\in S_{\text{in}}} \left\| H(x) - E\left(H(x)\right) \right\|_1 \tag{3.3}$$

该公式表明，希望水印嵌入前后图像之间的差距越小越好。此外，还可以选择性添加感知损失[64]和鉴别损失[56]，以进一步提高图像质量。

更重要的是，为了保证水印系统在频域上的高保真，必须针对不同的频率分量进行优化。即，希望含水印图像的频域特征尽可能接近于添加水印前图像的频域特征。因此，本章需要优化公式(3.4):

$$\mathcal{L}_4 = \sum_{k=1}^{6}\sum_{x\in S_{\text{in}}} \frac{\left\| \text{DWT}_k\left(H(x)\right) - \text{DWT}_k\left(E\left(H(x)\right)\right) \right\|_1}{|N|} \tag{3.4}$$

其中$\mathrm{DWT}_k(\cdot)$表示分频层得到的不同频率分量（即$HL, LH, HH, HL_2, LH_2, HH_2$）。通过优化公式(3.4)，嵌入网络$E$所标记的含水印图像的不同频域分量与其对应的不含水印图像的不同频域分量的差距将尽可能的小。这样，通过优化公式(3.3)和公式(3.4)就保证了水印嵌入前后，图像在空间域和频域上的变化尽可能的小，从而实现了水印系统的高保真。

最终，嵌入网络E、鉴别器D、提取网络R的总损失函数为公式(3.5):

$$L = \mathcal{L}_{ext} + \beta_1 \mathcal{L}_3 + \beta_2 \mathcal{L}_4 \tag{3.5}$$

其中β_1和β_2是可调超参数，嵌入网络和提取网络将进行联合训练。

3.3 实验结果与分析

3.3.1 实验设置

数据集：为了证明本章方法的有效性，本节评估了它在两个任务上的表现，即 paint transfer[56]和 de-raining[61]。如图 3.6 所示，paint transfer 任务输入人物轮廓图像，神经网络将学习上色技巧并补充颜色，最终输出上色完毕的彩色图像；de-raining 任务输入有雨图像，神经网络将学习去雨技巧，最终输出去雨完毕的图像。

图 3.6 数据集示例

对于 paint transfer 任务，本节采用了广泛使用的 Danbooru2019 数据集[60]。从该数据集中随机选择 15000 幅图像，其中 8000 幅图像用于阶段 A 的训练，遵循了文献[56]中描述的训练策略。在阶段 B 的训练中，剩下的 7000 幅图像被分成三个不相交的子集：6000 幅用于训练，500 幅用于验证，500 幅用于测试。对于 de-raining 任务，本节从 de-raining 数据集[61]中随机选择了 2000 幅图像。其中，1000 幅图像被用于阶段 A 的训练，遵循了文献[61]中描述的训练策略。在阶段 B 的训练中，剩下的 1000 幅图像被分成三个不相交的子集：800 幅用于训练，100 幅用于验证，100 幅用于测试。所有图像默认设置为$256\times256\times3$大小。本节分别使用了 Lena 和 IEEE 作为水印。需要注意的是，为了保证水印提取网络的输出维度匹配，水印大小被设置为$64\times64\times3$，这是因为经过两级 DWT 后，图像的低频分量大小变成了原始大小的$1/4$。

超参数设定：对于可调超参数，本章经验性的设置$\alpha=0.5$，$\beta_1=\beta_2=1$。使用 Adam 优化器进行训练，学习率设置为2.0×10^{-4}。本章的方法是在单个 RTX 3090 GPU 上执行的，该 GPU 使用 CuDNN 和 CUDA 为深度神经网络提供加速。

评价指标：两个常用的评价指标被用来评估含水印图像的质量，即峰值信噪比（Peak Signal-To-Noise Ratio, PSNR）[65]以及结构相似性（Structural Similarity, SSIM）[65-66]。对于每个度量，值越高，质量越好。误码率（Bit Error Ratio, BER）是衡量二值水印重建质量的指标。误码率越低，提取效果越好。此外，为了评估水印提取的性能，本章定义了一个新的度量，称为成功率（Success Rate, SR）。SR 确定为成功提取水印的百分比。如果提取水印的 PSNR 大于 35 dB 或 BER 小于0.5×10^{-3}，则认为水印提取成功。最后，利用 DCT 频域检测结果来衡量模型水印在频域上的保真度。

3.3.2 定性和定量实验结果

本节将提供本方法的定性和定量实验结果。为验证所提方法的泛化能力，实验在两个不同的数据集上进行。此外，实验中嵌入了两种类型的水印：彩色图像

水印（Lena）和二进制图像水印（IEEE）。这两种水印在视觉上均不可见，需通过水印提取网络进行提取和检测。彩色图像水印包含了图像每个像素的 RGB 分量，而二进制图像水印则是数字化图像，其每个像素只有两种可能的值。

空间域保真度：图 3.7 和图 3.8 提供了可视化示例。其中，图 3.7 演示了 paint transfer 任务的实验结果；图 3.7(a)表示主机网络的输出，即不含水印的图像；3.7(b)表示嵌入网络的输出，即含水印图像；图 3.7(c)表示含水印图像与不含水印图像之间的残差分析；图 3.7(d)表示对应的 DCT 频域检测结果；图 3.7(e)表示从含水印图像中所提取的水印。很容易观察到，图 3.7(a)展示的不含水印图像与图 3.7(b)展示的含水印图像之间非常相似，这表明本章提出的框架不仅保证了含水印图像良好的视觉效果，还避免在空间域引入明显的伪影。并且，图 3.7(c)所展示的残差分析进一步表明，残差中无法获取任何的水印信息，这意味着水印并不是简单的添加到图像上，而是以一种不可见的方式嵌入。此外，图 3.7(e)所示的从含水印图像中提取的水印图像具备令人满意的视觉效果，这表明所提出的框架可以实现可靠的所有权验证。类似的，图 3.8 展示了所提出算法在保护 de-raining 任务时的实验结果，实验结果与图 3.7 所展示的结果类似。

为了进一步说明本章方法的优越性，本章使用了 PSNR 和 SSIM 指标来评估含水印图像的视觉质量。具体而言，计算含水印图像与不含水印图像之间的 PSNR

图 3.7 paint transfer 任务的示例

图 3.8 de-raining 任务的示例

和 SSIM，值越高说明嵌入水印前后的图像变化越小，实现了越高的保真度和隐蔽性。另外，本章使用 PSNR（针对彩色水印）和 BER（针对二值水印）来评估提取水印的质量，水印的质量越高，水印系统的可靠性越好。实验结果见表 3.1 和表 3.2。可以发现，含水印的图像和提取水印都具有较高的 PSNR 和 SSIM，这表明含水印的图像和提取的水印都具有令人满意的图像质量，所提出的模型水印框架在空间域中达到了令人满意的保真度和水印提取性能。值得一提的是，所有情况下水印提取的成功率 SR 都是 100%，这表明提取水印能够被提取出来，并能可靠的用于目标 DNN 模型的知识产权验证和溯源。

表 3.1 嵌入彩色水印时含水印图像和提取水印的质量评价

任务	嵌入水印	PSNR	SSIM	水印 PSNR	SR
paint transfer	Lena	40.55 dB	0.988	64.06 dB	100%
de-raining	Lena	42.69 dB	0.991	60.86 dB	100%

表 3.2 嵌入二值水印时含水印图像和提取水印的质量评价

任务	嵌入水印	PSNR	SSIM	水印 BER	SR
paint transfer	IEEE	41.24 dB	0.990	0	100%
de-raining	IEEE	42.88 dB	0.992	0	100%

频域保真度：注意，本章工作最重要目标是抑制生成式模型水印中的高频伪影。为此，本章分析了测试（含水印）图像的频域特征，图 3.7(d)和图 3.8(d)所示的 DCT 热图展示了评估的结果。通过观察发现，图 3.7(d)和图 3.8(d)展示的 DCT 频域检测结果中没有明显的格纹状高频伪影，并且与 2.1.3 节中所展示的不含水印图像的 DCT 热图十分相似。这意味着本章所提出的生成式模型水印框架有效地减轻了高频伪影，使得含水印图像的频域检测结果接近于不含水印图像的频域检测结果，从而确保了水印系统在频域上的高保真。

上述基于图像空间域和频域的实验结果都表明，本章提出的水印框架比之前的模型水印技术[43-44]有显著改进，成功抑制了高频伪影。可以发现，含水印图像的 DCT 热图中不存在任何的高频伪影，同时检测结果与自然图像的结果十分相似。综上所述，所提出的框架实现了高保真的生成式模型水印技术，并进一步提升了水印系统的隐蔽性和安全性。

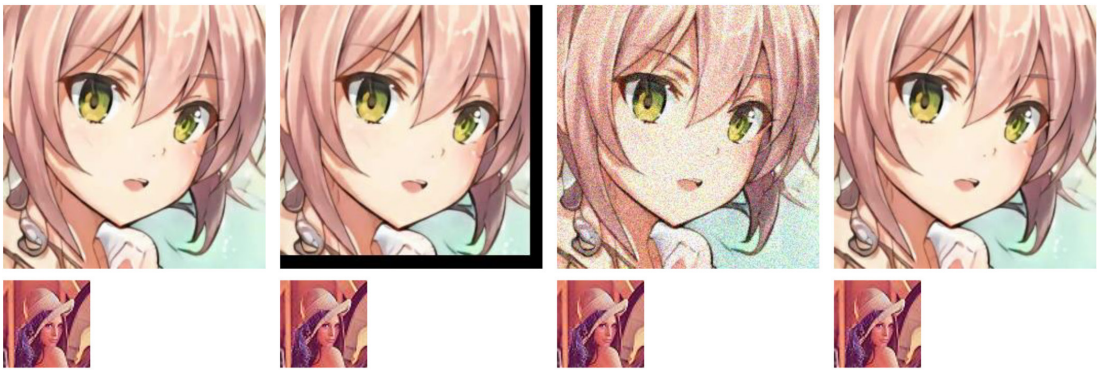
3.3.3 对预处理攻击的鲁棒性

稳健的数字水印系统应能抵御现实场景中可能出现的攻击。本节将重点放在抵抗预处理攻击上。预处理攻击是指攻击者对图像进行图像编辑操作（例如，添加噪声，裁剪图像等），从而阻碍水印的提取。为了提升水印系统对预处理攻击的鲁棒性，本节采用了对抗训练的方法。具体而言，在模型水印训练过程中，本章所提出的方法会对嵌入网络所标记的图像执行各种预处理攻击操作，受攻击的样本图像会继续被送入后续的频率分离层和水印提取网络，以进行对抗训练，从而增强水印提取能力。

由于计算资源有限，本节只考虑了三种常见的预处理攻击：噪声添加攻击、大小调整攻击和裁剪攻击。对于噪声添加攻击，本节在训练过程中将 $u=0$ 和 $\sigma\in(0,0.2)$ 的高斯噪声应用于每幅图像。对于大小调整攻击，本章将输入图像随机调整到 $\left[128^2,512^2\right]$ 范围内较小或较大的尺寸。对于裁剪攻击，本章只保留了图像在 $\left[192^2,256^2\right]$ 范围内的像素，其余像素被设置为 0 以模拟裁剪。值得注意的

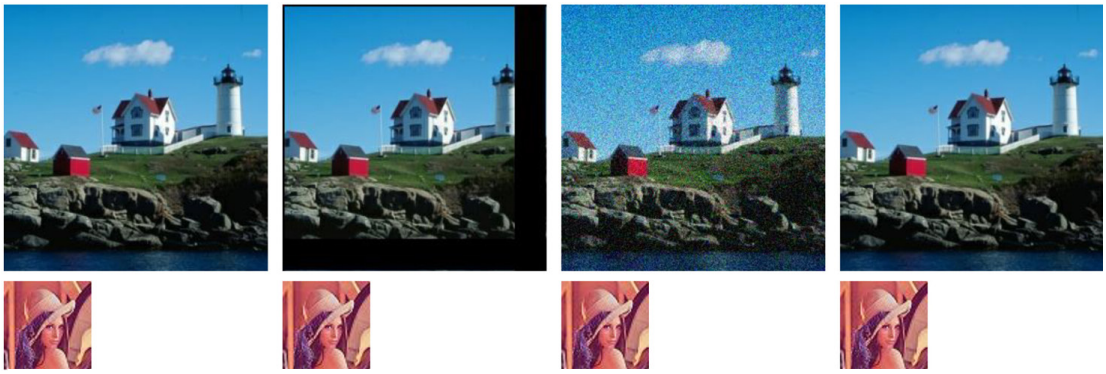
是，虽然本章只考虑了有限的攻击手段，但只要有足够的计算资源，仍有进一步探索的空间。

图 3.9 和图 3.10 分别提供了在保护 paint transfer 任务和 de-raining 任务时遭受预处理攻击的含水印图像和相应提取水印的示例。图 3.9(a)展示了未遭受攻击的含水印图像和其对应的提取水印；图 3.9(b)展示了遭受裁剪攻击后的含水印图像和其对应的提取水印；图 3.9(c)展示了遭受加噪攻击后的含水印图像和其对应的提取水印；图 3.9(d)展示了遭受调整大小攻击后的含水印图像和其对应的提取水印。图 3.10 展示了类似的结果。从这些视觉例子可以看出，尽管经过了不同的预处理操作，含水印的图像遭受了一定程度的破坏，但是嵌入的水印仍然可以很好地提取出来。这表明对抗性训练有效地提高了水印系统对预处理攻击的鲁棒性。



(a)无攻击　(b)裁剪攻击　(c)加噪攻击　(d)调整大小攻击

图 3.9 paint transfer 任务遭受预处理攻击时的示例



(a)无攻击　(b)裁剪攻击　(c)加噪攻击　(d)调整大小攻击

图 3.10 de-raining 任务遭受预处理攻击时的示例

定量结果如表 3.3 至表 3.8 所示。可以发现，对于同一类攻击，随着攻击强度的增加，提取水印的 PSNR 开始减小，误码率开始增大，提取成功率开始减小。例如，在遭受最大强度的大小调整攻击后，本章所提出的方法（保护 de-raining 任务时，嵌入 Lena 水印）所提取水印的质量从 60.86 dB（未遭受任何攻击）下降到 44.79dB（调整大小至128^2），水印提取的成功率从 100%下降到 76.40%。但总的来说，水印的提取仍然保证了高质量和高成功率，所提取水印的质量至少能保持在 40 dB/0.001 以上，水印提取成功率也至少能保持到 75%左右。

表 3.3 噪声添加攻击下提取水印的质量结果

任务	嵌入水印	$\sigma = 0.1$	$\sigma = 0.15$	$\sigma = 0.2$
paint transfer	Lena	58.26 dB	56.33 dB	54.28 dB
paint transfer	IEEE	0.0002	0.0005	0.0008
de-raining	Lena	57.18 dB	54.44 dB	52.88 dB
de-raining	IEEE	0.0002	0.0007	0.0008

表 3.4 大小调整攻击下提取水印的质量结果

任务	嵌入水印	$128^2 \times 3$	$196^2 \times 3$	$512^2 \times 3$
paint transfer	Lena	59.46 dB	62.70 dB	62.98 dB
paint transfer	IEEE	0.0013	0.0004	0.0002
de-raining	Lena	44.79 dB	56.91 dB	58.57 dB
de-raining	IEEE	0.0009	0.0006	0.0006

表 3.5 裁剪攻击下提取水印的质量结果

任务	嵌入水印	16	32	64
paint transfer	Lena	49.43 dB	46.94 dB	43.05 dB
paint transfer	IEEE	0	0.0002	0.0004
de-raining	Lena	59.49 dB	58.58 dB	53.98 dB
de-raining	IEEE	0	0.0001	0.0002

表 3.6 噪声添加攻击下水印提取的成功率

任务	嵌入水印	$\sigma=0.1$	$\sigma=0.15$	$\sigma=0.2$
paint transfer	Lena	99.80%	98.60%	97.50%
paint transfer	IEEE	99.60%	98.40%	97.80%
de-raining	Lena	99.50%	98.80%	96.30%
de-raining	IEEE	99.70%	98.30%	97.60%

表 3.7 大小调整攻击下水印提取的成功率

任务	嵌入水印	$128^2\times3$	$196^2\times3$	$512^2\times3$
paint transfer	Lena	79.50%	96.30%	98.80%
paint transfer	IEEE	81.60%	97.60%	98.90%
de-raining	Lena	76.40%	96.20%	97.10%
de-raining	IEEE	86.80%	97.20%	98.30%

表 3.8 裁剪攻击下水印提取的成功率

任务	嵌入水印	16	32	64
paint transfer	Lena	100%	96.60%	95.60%
paint transfer	IEEE	100%	95.50%	94.50%
de-raining	Lena	100%	96.30%	93.70%
de-raining	IEEE	100%	96.70%	94.80%

3.3.4 对高频信息移除攻击的鲁棒性

本节将所提出的方法与文献[43]、文献[44]和文献[46]提出的方法进行了比较。如果对手事先知道频域相关知识，他/她可能会过滤掉含水印图像的高频成分，以此去除嵌入的水印。并且，高频信息的滤除并不会严重的降低图像质量。为了模拟这种攻击，本节过滤掉了含水印图像的高频成分。具体而言，给定一幅含水印图像，确定该图像每个通道的 DCT 系数。然后，保持每个通道左上角196×196范围内的 DCT 系数不变，然后将其他区域的系数改为零值。

实验结果如表 3.9 所示。可以发现，对于所有对比方法，滤波后彩色水印的平均 PSNR 显著下降，二进制水印的平均误码率显著增加。实验结果表明，先前的方法[43, 44, 46]在遭受高频信息滤除攻击后提取的水印质量差。相比之下，对于本章所提出的方法来说，滤波前提取水印的平均 PSNR/BER 与滤波后的平均 PSNR/BER 之间的差异非常小，这表明所提出的方法在遭受高频信息滤除攻击后仍然可以提取出质量较高的水印。

表 3.9 滤除标记图像高频分量前后提取水印的质量，上标*表示应用过滤操作

方法	水印	文献[43]	文献[44]	文献[46]	本章方法
paint transfer	Lena	27.25 dB	35.18 dB	29.62 dB	64.06 dB
paint transfer*	Lena	17.38 dB	12.51 dB	10.16 dB	60.33dB
paint transfer	IEEE	0.0033	0.0026	0.0014	0
paint transfer*	IEEE	0.4162	0.5259	0.5733	0.0002
de-raining	Lena	27.50 dB	38.61 dB	26.02 dB	60.86 dB
de-raining*	Lena	12.61 dB	12.67 dB	14.13 dB	59.59 dB
de-raining	IEEE	0.0015	0.0002	0.0003	0
de-raining*	IEEE	0.5623	0.4442	0.4517	0.0006

因此，可以推断，相比于相关研究，本章提出的框架具有更强的抗高频信息移除攻击能力。这一优势源自于本章方法施加的特定约束，即通过小波变换分离层获取图像不同频域成分和联合损失函数优化，精确限制水印被嵌入到样本的低频区域中。此策略确保水印不会受到高频滤波攻击的影响，因为这类攻击不涉及图像的低频成分。鉴于低频成分包含了图像的主要信息，攻击者在尝试滤除低频成分的同时难以维持图像质量，这使得针对低频分量的攻击变得不可取。相比之下，其他方法往往忽视了频域内的保真度，它们在整个频谱中无差别地嵌入水印，这使得攻击者在过滤含水印图像的高频信息时会破坏水印信息的完整性，从而影响版权保护的有效性。

3.3.5 与现有方法比较

本节进一步将本章提出的方法与文献[43]、文献[44]和文献[46]进行比较。表3.10 给出了不同方法在空间域/频域上对预处理攻击的鲁棒性和保真度的比较结果。在表 3.10 中，“是”表示对应的方法对相应的攻击具有鲁棒性，或者在相应的域中实现了高保真。“部分”表示对应的方法的一部分在相应的域内具有较好的保真度，即另一部分图像引入了伪影。“否”则表示对应的方法不具有鲁棒性，或者在相应的域中未能实现高保真。可以发现，现有的方法只能达到部分目的，而本章提出的方法可以达到全部目的。即，本章提出的方法具有抗预处理攻击的能力，同时在空间域和频域都实现了良好的保真度。

表 3.10 不同模型水印方法的鲁棒性和保真度

方法	对预处理攻击的鲁棒性	空间域保真度	频域保真度
文献[43]	是	是	否
文献[44]	否	部分	否
文献[46]	是	部分	否
本章方法	是	是	是

3.4 本章小结

随着图像处理模型在物联网中的数据处理、数据增强及隐私保护等任务中变得日益重要，确保这些模型的版权得到保护显得尤为紧迫。然而，现有的生成式模型水印技术在水印嵌入过程中可能无意地引入了高频伪影，从而增加了攻击者检测并移除水印的风险，对知识产权保护构成了严重威胁。

为了应对这一挑战，本章提出了一种基于离散小波变换的生成式模型水印框架，该框架能够有效地缓解高频伪影问题，提高水印系统的保真度。框架的核心思想是通过优化输出图像中水印的分布来抑制高频伪影。具体而言，该框架通过频域分离层将嵌入网络所标记图像的频谱分解为不同频率区域，然后再通过联合

训练和损失函数的优化，使得水印嵌入到输出图像的低频区域。综合实验结果表明，本章提出的水印技术抑制了高频伪影，具有很高的保真度，并能抵御常见的预处理攻击。

由于计算资源的限制，本章仅考察了几种常见的预处理攻击。必须坦诚地承认，面对现实世界中多样化的攻击，本章提出的框架无法保证绝对的鲁棒性。这是因为预测并应对所有潜在对手的攻击是不现实的，而某些攻击可能会针对特定方法的弱点。实际上，大多数现有方法也仅在特定攻击下展现出较强的防御能力。然而，通过减少水印系统中的高频伪影，极大地提高了水印系统的保真度和隐蔽性。高保真度和隐蔽性显著降低了水印遭受攻击的风险，从而提升了水印系统版权保护的安全性和可靠性。

在本章中，所提出的方法通过联合训练和损失函数优化过程将水印嵌入到生成图像的低频区，以减少水印嵌入过程导致的高频伪影。然而，如第 2.3 节所分析，高频伪影的产生还可能源于水印嵌入网络本身的采样结构。本章所提出的方法未能充分考虑此因素导致的高频伪影。因此，为进一步提高生成式模型水印的保真度，本节建议未来的研究可以考虑解决嵌入网络结构所导致的高频伪影问题。据此，下一章所提出的方法将更为全面，同时考虑了嵌入网络和水印嵌入过程导致的高频伪影。

第四章 基于频域扰动的生成式模型水印

4.1 研究动机

本章研究了一系列旨在保护图像生成网络的技术[43-46]，这些技术借鉴了深度隐写术中的先进技术[51, 67]，即都采用了嵌入网络-提取网络的架构。可以发现，生成式模型水印的核心正是利用水印嵌入网络将水印嵌入到被保护模型的输出中，并通过水印提取网络来提取水印，从而实现所有权的验证。水印嵌入网络及水印添加过程起到了至关重要的作用。然而本文第 2.3 节对高频伪影的成因分析中指出，当前生成式模型水印所面临的高频伪影问题正是由水印嵌入网络及水印添加过程所导致的。

基于此，现有的生成式模型水印普遍面临高频伪影问题，这降低了水印系统的保真度。为了解决生成式模型水印中的高频伪影问题，本文的第三章提出了一种生成式模型水印技术，该技术旨在优化水印在输出图像中的分布。具体而言，该技术通过应用小波变换分离层将水印嵌入至图像低频区域，从而有效减少了水印嵌入过程中产生的高频伪影。然而，此方法并未解决由水印嵌入网络固有结构所引起的高频伪影问题。为此，文献[70]提出了一种基于抗混叠技术的水印嵌入网络方案，有效地抑制了由水印嵌入网络及添加水印过程所引起的高频伪影问题。但是，由于该方法依赖于低通滤波技术，这可能导致含水印图像在高频信息上的损失，进而影响了图像的质量。

综上所述，尽管第三章所提出的方法和文献[70]的方法在一定程度上解决了生成式模型水印在频域上的高频伪影问题，但仍存在一定的局限性。为了实现更加有效的生成式模型水印技术，本章全面考虑了水印嵌入网络和水印添加过程所导致的高频伪影问题。

简而言之，本章的目的是引入一项消除高频伪影的高保真生成式模型水印技术。该技术致力于消除高频伪影，并在水印的隐蔽性和有效性之间找到平衡点，从而提供一种更优化的生成式模型水印解决方案。

4.2 方案设计

4.2.1 框架概述

正如前文所述，现有的生成式模型水印技术在频域上的高频伪影问题显著影响了水印系统的保真度。为了应对这一挑战，本章提出了一种高保真的生成式模型水印框架。现有的生成式模型水印技术[43-46]普遍采用水印嵌入网络-水印提取网络的框架，这种框架不可避免地导致高频伪影的产生。与之不同，本章提出的方法采用了频域扰动生成网络-水印提取网络的新框架。这一框架的关键在于不使用传统的水印嵌入网络实现水印的嵌入，从而规避了由嵌入网络本身可能引入的高频伪影。此外，频域扰动生成网络会产生频率扰动，该扰动随后被加入到载体图像的低频部分中，这不仅完成了水印的嵌入，同时也限制了水印在图像中的分布，以优化其隐蔽性和抗干扰能力。在此基础上，框架进一步集成了水印提取网络，以从含水印图像的低频区域中准确地提取水印。

如图 4.1 所示，展示了基于频域扰动的生成式模型水印框架。给定一幅尺寸为 256×256×3 的载体图像（由待保护的主机网络生成），其频域中心128×128×3大小范围被定义为低频区域，是水印嵌入的目标位置。为了完成水印的嵌入，频

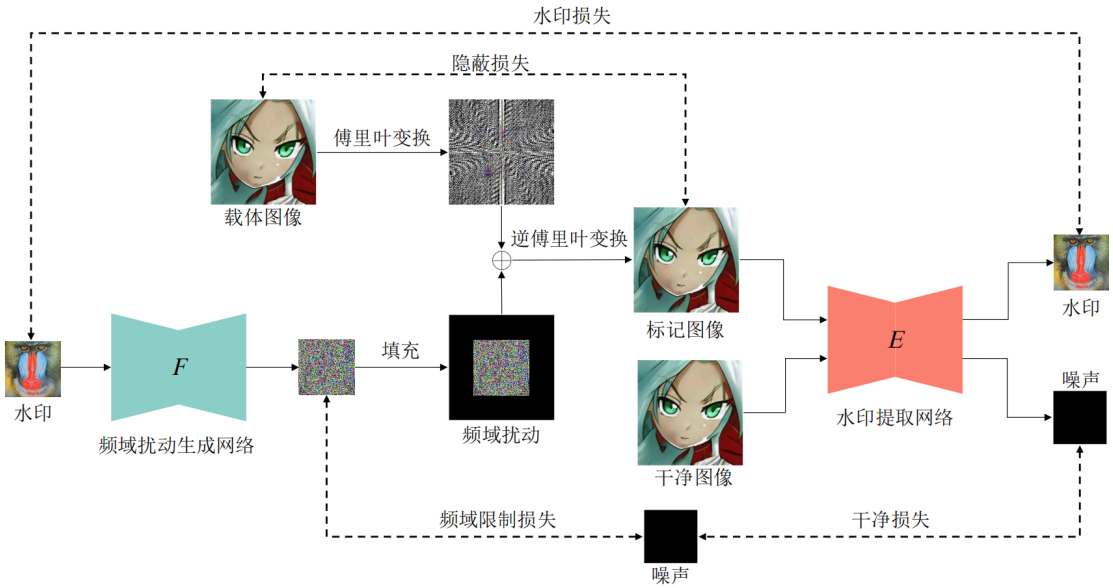


图 4.1 基于频域扰动的生成式模型水印框架图

域扰动生成网络 F 需要生成频域扰动，并将其添加到载体图像的低频区域中。具体而言，一个大小为128×128×3的固定水印被输入到频域扰动生成网络，该网络随后输出一个大小为128×128×3的频域扰动。频域扰动将被进一步填充到256×256×3的尺寸，以匹配载体图像的尺寸。值得注意的是，引入的频域扰动仅影响载体图像的低频区域，而不影响其高频区域。最终，将频域扰动直接添加到载体图像的频域即实现了水印嵌入。

基于此，本章的方法能有效地将频域扰动嵌入到载体图像的低频分量中，同时避免嵌入网络的使用。通过绕过水印嵌入网络，本章提出的方法能确保载体图像不受深度网络固有结构所导致的高频伪影影响；通过控制频域扰动的嵌入范围，本章所提出的方法可以选择性地修改载体图像的低频区域，而高频区域则保持不变，这种有针对性的策略有效地消除了水印嵌入过程中常见的高频伪影。最后，通过联合训练，水印提取网络就能从标记图像中准确提取水印，从而实现了知识产权的保护。

4.2.2 结构设计

在本章中，频域扰动生成网络 F 采用了类似 U-Net 的网络结构[6]，输入和输出维度设置为128×128×3。对于水印提取网络 E，本章采用了 Cycle-GAN[62]中的类似 ResNet 的网络结构，输入和输出维度分别为256×256×3和128×128×3。载体图像是由受保护模型生成的图像，大小为256×256×3。频域扰动是由频域扰动生成网络生成的，大小为128×128×3。频域扰动将被添加到载体图像的频域中以实现水印的嵌入。为了使频域扰动与载体图像频域尺寸匹配，需要对频域扰动进行填充（填充区域设置为0）。本章将频域扰动直接加入载体图像的频域中（执行快速傅立叶变换以获取图像的频域），然后执行反快速傅立叶变换将其转换回空间域，从而得到含水印图像。重要的是，所添加的频域扰动只影响了载体图像的低频区域，没有影响高频区域。

同样，本章的目标是在损失函数的限制下，让含水印图像与载体图像尽可能相似，使得频率扰动对含水印图像的影响降至最低。将含水印图像输入水印提取网络 E，期望水印提取网络 E 能学会如何从含水印图像的低频中提取水印。为防止水印提取网络 E 过度拟合，还会为其提供一些不含水印的负样本。为此，需要根据上述目标精心设计损失函数。最终，上述目标将通过联合训练水印提取网络和频域扰动生成网络来实现。

4.2.3 损失函数

在本生成式模型水印框架中，频域扰动生成网络 F 负责生成频率扰动 f，随后将该扰动添加到载体图像 C 的频域中，以得到含水印图像 C'。含水印图像 C' 随后被输入到水印提取网络 E，以便从中提取水印信息。值得注意的是，当水印提取网络 E 接收到未经修改的干净图像时，应输出噪声。更重要的是，本研究希望水印扰动 f 对含水印图像 C' 在图像空间域及频域上的影响最小化，即载体图像 C 与含水印图像 C' 在空间域和频域上尽可能相似。为了实现上述目标，本节设计了四个关键的损失函数：频域限制损失函数、水印损失函数、干净损失函数和隐蔽损失函数。这些损失函数在优化水印处理过程和确保其有效性方面发挥着关键作用。

频域限制损失函数：正如前文所述，为了最大程度地减小水印扰动对载体图像的影响，本研究旨在尽量优化扰动 f。即，目标是是将频率扰动 f 最小化。为了达成此目标，本章设计了频域限制损失函数，即公式(4.1):

$$\mathcal{L}_1 = \Sigma \left\| f \right\|_1 \tag{4.1}$$

其中，f 表示由频率扰动网络 F 产生的扰动，$f = F\left(wm\right)$。$F\left(\cdot\right)$ 表示频率扰动网络 F 的输出，wm 表示固定水印。

水印损失函数：本章在载体图像C的频域内直接加入频率扰动f来实现水印任务。因此，得到的含水印图像C'可表示为：$C' = IFFT\left(f \cdot \alpha + FFT\left(C\right)\right)$。这里，α代表一个可调节的超参数，$FFT\left(\cdot\right)$代表快速傅里叶变换，而$IFFT\left(\cdot\right)$代表反快速傅里叶变换。通过将频率扰动$f$作为水印加入到载体图像中，本研究获得了含水印的图像C'。更为关键的是，水印的提取依赖于水印提取网络E，它应该能从含水印图像C'中准确提取出水印。基于此目的，本章进一步设计了水印损失函数，具体如公式(4.2)所述：

$$\mathcal{L}_2 = \frac{1}{|N|}\sum_{x \in C'} \left\| E\left(x\right) - wm \right\|_1 \tag{4.2}$$

上式中，x表示属于C'的含水印图像。N表示像素总量。$E\left(\cdot\right)$表示水印提取网络的输出。通过优化公式(4.2)可以使得水印提取网络E从含水印图像C'中准确提取出水印。

干净损失函数：另一方面，为确保水印提取网络E的可靠性及防止过拟合现象，水印提取网络E不应该从不含水印的图像中提取出任何水印信息，而是应从这些图像中提取出噪声。基于此理念，本节引入一个干净损失项，即公式(4.3)：

$$\mathcal{L}_3 = \frac{1}{|N|}\sum_{x \in C} \left\| E\left(x\right) - noise \right\|_1 \tag{4.3}$$

其中x属于干净的载体图像C。通过优化公式(4.3)，水印提取网络能从干净的图像中提取噪声。在本章中，噪声默认为零值。

隐蔽损失函数：最后，为确保含水印图像具备最优的视觉质量，并以空间上不可察觉的方式进行水印嵌入，本章力求最大限度地缩小载体图像与含水印图像之间的差异。基于这一目标，本章特别设计了隐蔽损失函数，如公式(4.4)所示：

$$\mathcal{L}_4 = \frac{1}{|N|}\Sigma \left\| C' - C \right\|_1 \tag{4.4}$$

除非另有说明，否则本章默认使用 L_1 范数作为距离度量。最后，联合训练频域扰动生成网络 F 和水印提取网络 E，总损失函数如下公式(4.5)所示：

$$\mathcal{L}_{total} = \beta_1\mathcal{L}_1 + \beta_2\mathcal{L}_2 + \beta_3\mathcal{L}_3 + \beta_4\mathcal{L}_4 \tag{4.5}$$

其中，β_1，β_2，β_3，β_4 为可调节的超参数。

4.3 实验结果与分析

4.3.1 实验设置

数据集：为了验证所提方法的有效性，本节同样在两项不同的图像任务上进行了评估，即 paint transfer[56]和 de-raining[61]。在第 3.3.1 节中已经对数据集进行了展示。对 paint transfer 模型进行训练后，本节随机抽取了由该模型生成的 4000 幅图像（即载体图像），其中 3000 幅用于生成式模型水印训练，500 幅用于验证，500 幅用于测试。在训练好 de-raining 模型后，本节随机选择了由该网络生成的 2000 幅图像（即载体图像），其中 1800 幅用于生成式模型水印训练，100 幅用于验证，100 幅用于测试。本节分别使用了 Baboon 和 MDPI 作为水印图像，水印大小固定为 $128\times128\times3$。

超参数设定：在可调参数方面，本节经验性的设置了 $\alpha = 5\times10^5$，$\beta_1 = \beta_3 = \beta_4 = 1$，$\beta_2 = 5$。本节使用了 Adam 优化器进行训练，学习率为 2.0×10^{-4}。本章的方法是在带有 CuDNN 加速功能的 TITAN RTX GPU 上实现的。

评价指标：评估含水印图像质量的常用指标有两个，即 PSNR 和 SSIM。评估二值水印质量的指标采用了 BER。此外，为了评估水印提取的性能，本节重新定义了水印提取成功率 SR，如果相应的 PSNR 大于 25 dB 或相应的 BER 小于 1.0×10^{-2}，则认为水印提取成功。最后，利用 DCT 检测结果来衡量水印系统在频域上的保真度。另外，本章定义了频域隐蔽性能（Frequency Hiding Performance, FP），即计算了含水印图像的 DCT 检测结果与不含水印图像的 DCT 检

测结果之间的学习感知图像补丁相似度（Learned Perceptual Image Patch Similarity, LPIPS）[68]，该指标值越小越好。

4.3.2 定性和定量实验结果

在本节中，将展示本章所提出的生成式模型水印算法的定性与定量实验结果。本研究的核心目标在于消除高频伪影对生成式模型水印的干扰，重点提高水印系统的保真度和隐蔽性。值得指出的是，出色的隐蔽性进一步增强了水印系统的安全性。为达成上述目的，本章提出了一个创新性框架，旨在解决生成式模型水印所面临的高频伪影问题。

空间域保真度：图 4.2 为本章所提出方法的可视化结果提供了直观展示。图 4.2(a)(b)(c)展示了保护 paint transfer 模型的实验结果，图 4.2(d)(e)(f)则展示了保护 de-raining 模型的实验结果。具体而言，图 4.2(a)(d)呈现了主机网络的输出，即未嵌入水印的载体图像；图 4.2(b)(e)展示了本章提出的生成式模型水印的输出，即含水印图像；而图 4.2(c)(f)则显示了从含水印图像中提取的水印。这些结果清楚地表明，在空间域内，嵌入水印后的图像与原始载体图之间几乎无可见视觉差异，这强调了本章所提的方法能在不牺牲图像质量的前提下，有效的完成水印的嵌入。本方法不仅取得了卓越的视觉效果，还确保了水印系统在空间域上极佳的保真度。同时，图 4.2(b)(e)所展示的水印图像也呈现出了令人满意的视觉效果。从含水印的图像中提取出的高质量水印对于可靠地验证所有权至关重要，优秀的水印提取质量进一步表明了水印系统的可靠性。

定量评估结果如表 4.1 和表 4.2 所示，量化了本章提出的生成式模型水印在图像空间域的性能。在本节中，采用了 PSNR 和 SSIM 来对含水印图像和提取水印的质量进行了评估。较高的 PSNR 和 SSIM 值指示了载体图像与含水印图像之间的高相似度，这表明本章提出的生成式模型水印在完成水印嵌入的过程中成功保持了含水印图像的高质量，从而确保了水印系统的高保真。另外，提取水印的质量评估采用了 PSNR 和 BER。表 4.1 和表 4.2 的评估指标显示，嵌入水印的图

图 4.2 paint transfer 和 de-raining 任务的示例

像质量（PSNR 大于 50 dB）和提取水印的质量（PSNR 大于 40 dB 或 BER 小于 0.001）均表现出色，水印提取成功率 SR 达到 100%。上述结果充分表明，本章提出的生成式模型水印有效地确保了空间域上良好的保真度。

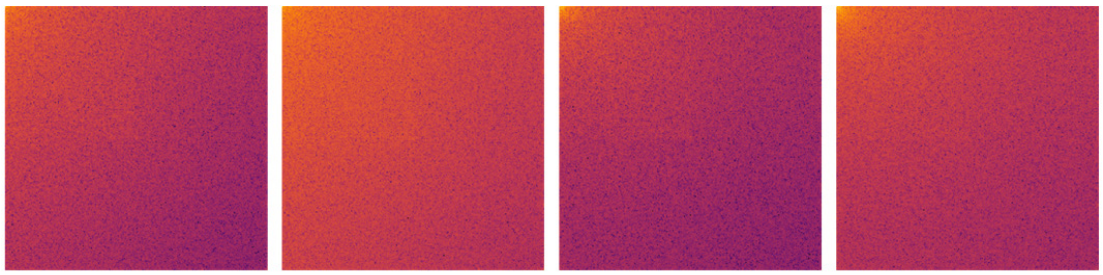
表 4.1 嵌入彩色水印时标记图像和提取水印的质量评估

方法	水印	PSNR	SSIM	水印 PSNR	SR
paint transfer	Baboon	50.58 dB	0.998	42.88 dB	100%
de-raining	Baboon	53.34 dB	0.999	46.47 dB	100%

表 4.2 嵌入二值水印时标记图像和提取水印的质量评估

方法	水印	PSNR	SSIM	水印 BER	SR
paint transfer	MDPI	57.18 dB	0.999	0.0009	100%
de-raining	MDPI	54.59 dB	0.999	0.0010	100%

频域保真度：更重要的是，本章的主要目标是通过消除高频伪影来提高生成式模型水印的保真度。为了说明本章提出的生成式模型水印技术在频域中良好的保真度，本节检查了由本章方法所生成含水印图像的 DCT 热图。如图 4.3 所示，图 4.3(a)表示保护 paint transfer 模型时嵌入 Baboon 的情况；图 4.3(b)表示保护 paint transfer 模型时嵌入 MDPI 的情况；图 4.3(c)表示保护 de-raining 模型时嵌入 Baboon 的情况；图 4.3(d)表示保护 de-raining 模型时嵌入 MDPI 的情况。可以发现，本章方法所生成的含水印图像的 DCT 热图与 2.1.3 节中所展示的不含水印图像的 DCT 热图非常接近。这表明本章提出的生成式模型水印技术所生成的图像非常自然，没有高频伪影。相比之下，以前的模型水印方法（如 3.1 节所示）会出现块状高频伪影。



(a)PT-Baboon (b)PT-MDPI (c)DE-Baboon (d)DE-MDPI

图 4.3 含水印图像的 DCT 检测结果

此外，为了定量评估本章提出的生成式模型水印在频域上的保真度，本节还评估了所提出方法的频域隐蔽性能 FP。结果如图 4.4 所示。其中，“-”前表示相应的图像任务，“-”后表示使用的水印；“DE”表示 de-raining，“PT”表示 paint transfer；“Wu”表示文献[43]的模型水印方法，“Zhang”表示文献[44]的模型水印方法，而“原始图像任务”表示仅执行图像处理任务的主机网络。很容易发现，本章提出的方法在各种任务中都有良好的表现。例如，在使用 MDPI 作为水印来保护 de-raining 主机网络的情况中，本章方法的频域隐蔽性能比文献[43]高出 44%。此外，本章提出的生成式模型水印方法的平均频域隐蔽性能比文献[43]的平均频域隐蔽性能最多提升了 24.9%，比文献[44]的频域隐蔽性能最多提升了 7.2%。上述基于图像空间域和频域的实验结果都表明，本章提出的生成式

模型水印方法所生成的含水印图像表面光滑，同时避免了频域上的高频伪影，这是一个显著的优势。以上结果突出表明了本章提出的框架在图像空间域和频域上都实现了很高的保真度。总而言之，基于这些实验的综合证据可以断言，本章提出的生成式模型水印方法有效地消除了高频伪影，在空间域和频域都实现了高保真，同时还保持了可靠的知识产权保护性能。

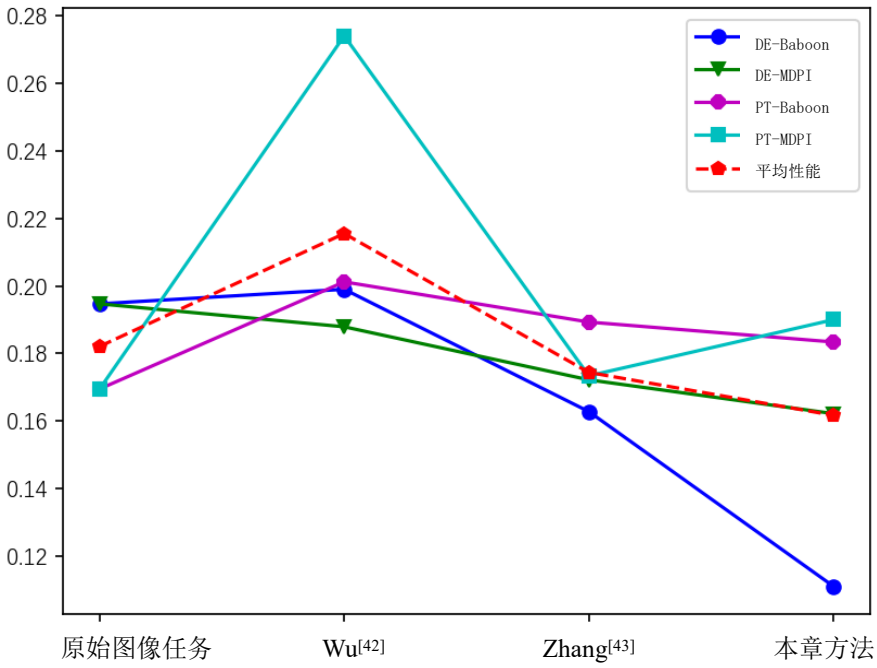


图 4.4 不同生成式模型水印方法在不同任务上的频域隐蔽性能

4.3.3 对预处理攻击的鲁棒性

稳健的模型水印系统必须能够应对现实世界中常见的攻击。本节同样重点考虑了对预处理攻击的鲁棒性。为了增强水印系统对这些攻击的鲁棒性，一种常见的策略是在训练集中加入预处理图像[43]。由于计算资源有限，本节的研究侧重于四种常见的预处理攻击：高频信息过滤攻击、调整大小攻击、裁剪攻击和翻转攻击。

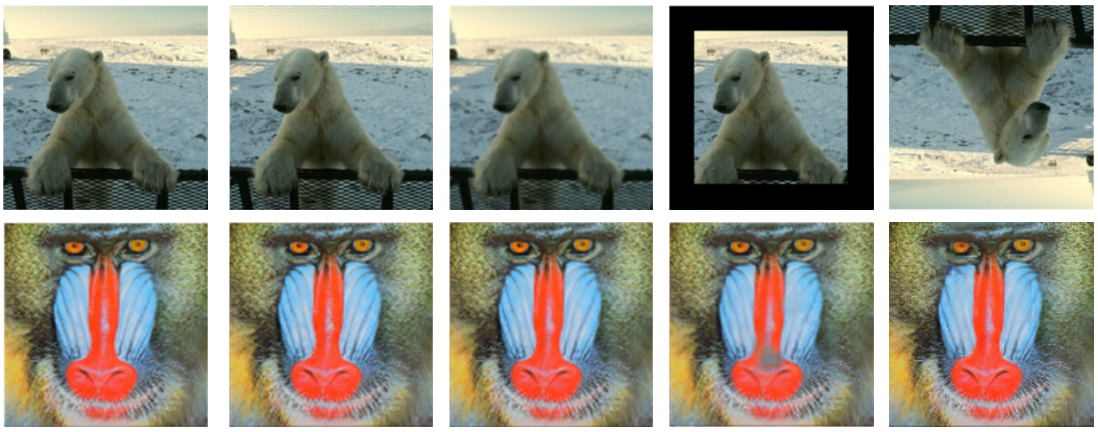
对于高频信息过滤攻击，事先掌握频域先验知识的攻击者可能会试图滤除含水印图像中的高频成分来消除水印。为了模拟这种攻击，含水印图像中的高频成分将被过滤。具体而言，本节计算了含水印图像每个通道的快速傅里叶系数，随

后只保留跨度为$\left[128^2,256^2\right]$范围的系数。对于调整大小攻击，攻击者可能会试图通过随机调整含水印图像的尺寸来消除嵌入其中的水印。为了模拟这种攻击，本节随机的改变输入图像尺寸，使其尺寸在$\left[128^2,512^2\right]$范围内变换。对于裁剪攻击，攻击者可能会通过裁剪含水印图像的部分信息来消除嵌入其中的水印。为了模拟裁剪攻击，本节保留了含水印图像$\left[192^2,256^2\right]$范围内的某些像素，然后再将其余像素设置为零。对于翻转攻击，攻击者可能会通过翻转含水印图像来消除水印。为了模拟翻转攻击，本节随机地对含水印图像进行了水平翻转或垂直翻转。同样值得注意的是，虽然本节中只考虑了上述 4 种常见的攻击，但如果有足够的计算资源，本章所提出的生成式模型水印还有进一步探索鲁棒性的潜力。

图 4.5 和图 4.6 展示了在保护 paint transfer 任务和 de-raining 任务时遭受预处理攻击的图像和相应提取水印的示例。图 4.5(a)展示了未遭受攻击的含水印图像和其对应的提取水印；图 4.5(b)展示了遭受高频信息滤除攻击后的含水印图像和其对应的提取水印；图 4.5(c)展示了遭受调整大小攻击后的含水印图像和其对应的提取水印；图 4.5(d)展示了遭受裁剪攻击后的含水印图像和其对应的提取水印；图 4.5(e)展示了遭受翻转攻击后的含水印图像和其对应的提取水印。图 4.6 展示了类似的结果。这些示例表明，尽管进行了各种预处理操作，嵌入的水印仍能以令人满意的质量被提取出来。

(a)无攻击　(b)高频滤除　(c)调整大小　(d)裁剪　(e)翻转

图 4.5 paint transfer 遭受预处理攻击时的示例



(a)无攻击　(b)高频滤除　(c)调整大小　(d)裁剪　(e)翻转

图 4.6 de-raining 遭受预处理攻击时的示例

进一步地，表 4.3 至表 4.10 展示了定量实验的结果。可以发现，虽然水印的提取质量随着攻击强度的升高而降低，但本章所提方法的水印提取质量依然能维持在较高水准。例如，在遭受最大强度的裁剪攻击后，本章所提出的方法（保护 paint transfer 任务时，嵌入 Baboon 水印）的提取水印的质量从 42.88 dB（未遭受任何攻击）下降到 30.87 dB（裁剪 64 ），水印提取的成功率从 100%下降到 79.50%。但是，在遭受各种类型攻击的不利条件下，提取水印的 PSNR/BER 依旧能保持在 25dB/0.1 左右，水印的提取成功率也能保持在 70%以上，这足以保证知识产权的认证。同样，当攻击强度减弱时，提取水印的质量也能迅速恢复。这表明对抗训练能有效的增强对预处理攻击的鲁棒性。总的来说，虽然水印提取质量会因攻击强度的增加而有所下降，但是本章所提出的生成式模型水印框架通常能保持较高的提取质量，从而实现可靠的版权认证。

表 4.3 高频信息过滤攻击下提取水印的质量

任务	嵌入水印	$128^2 \times 3$	$160^2 \times 3$	$192^2 \times 3$
paint transfer	Baboon	39.23 dB	40.88 dB	41.04 dB
paint transfer	MDPI	0.0025	0.0011	0.0010
de-raining	Baboon	44.85 dB	45.20 dB	45.52 dB
de-raining	MDPI	0.0016	0.0014	0.0013

表 4.4 大小调整攻击下提取水印的质量

任务	嵌入水印	$148^2 \times 3$	$196^2 \times 3$	$512^2 \times 3$
paint transfer	Baboon	37.93 dB	40.16 dB	30.87 dB
paint transfer	MDPI	0.0010	0.0011	0.0012
de-raining	Baboon	44.30 dB	45.59 dB	25.55 dB
de-raining	MDPI	0.0028	0.0013	0.0015

表 4.5 裁剪攻击下提取水印的质量

任务	嵌入水印	16	32	64
paint transfer	Baboon	37.08 dB	33.64 dB	30.87 dB
paint transfer	MDPI	0.0009	0.0009	0.0010
de-raining	Baboon	32.92 dB	29.20 dB	25.55 dB
de-raining	MDPI	0.0103	0.0110	0.0116

表 4.6 翻转攻击下提取水印的质量

任务	嵌入水印	水平	垂直
paint transfer	Baboon	28.85 dB	29.85 dB
paint transfer	MDPI	0.0017	0.0021
de-raining	Baboon	36.90 dB	36.12 dB
de-raining	MDPI	0.0026	0.0022

表 4.7 高频信息过滤攻击下水印提取的成功率

任务	嵌入水印	$148^2 \times 3$	$160^2 \times 3$	$192^2 \times 3$
paint transfer	Baboon	94.20%	97.40%	97.60%
paint transfer	MDPI	79.20%	87.00%	88.60%
de-raining	Baboon	94.00%	95.50%	97.00%
de-raining	MDPI	89.50%	90.00%	90.50%

表 4.8 大小调整攻击下水印提取的成功率

任务	嵌入水印	$128^2 \times 3$	$196^2 \times 3$	$512^2 \times 3$
paint transfer	Baboon	74.00%	94.00%	99.60%
paint transfer	MDPI	69.20%	80.20%	84.80%
de-raining	Baboon	72.00%	92.50%	96.00%
de-raining	MDPI	80.00%	83.00%	89.50%

表 4.9 裁剪攻击下水印提取的成功率

任务	嵌入水印	16	32	64
paint transfer	Baboon	98.80%	96.30%	79.50%
paint transfer	MDPI	98.90%	97.60%	81.60%
de-raining	Baboon	97.10%	96.20%	76.40%
de-raining	MDPI	98.30%	97.20%	86.80%

表 4.10 翻转攻击下水印提取的成功率

任务	嵌入水印	水平	垂直
paint transfer	Baboon	81.20%	83.00%
paint transfer	MDPI	88.80%	85.00%
de-raining	Baboon	75.50%	77.50%
de-raining	MDPI	88. 00%	84.50%

4.3.4 与传统图像水印方法的比较

本章提出了一种高保真的生成式模型水印技术，旨在保护图像生成网络的知识产权。本节希望强调本章所提出的方法与传统图像水印技术之间的区别，概括如下：

研究问题的新颖性：本章针对生成式模型水印中遇到的高频伪影问题，提出了一种旨在消除高频伪影并增强版权保护能力的生成式模型水印算法。该算法的主要目标是保障生成网络的版权安全，这与传统图像水印技术直接针对图像版权

保护的目的形成鲜明对比。当然，所提出的方法也能够很好地保护图像网络所生成图像的知识产权。

方案的创新性：本章所提出的方法采纳了端到端的深度学习策略。通过深度神经网络实现水印的嵌入与提取。在细致的设计下，本方法所提出的生成式模型水印技术成功解决了高频伪影问题，进而提高了水印系统的保真度。相较于此，传统图像水印技术的开发往往依赖于对各类数学技术的深入理解，并需要精心设计嵌入与提取过程。

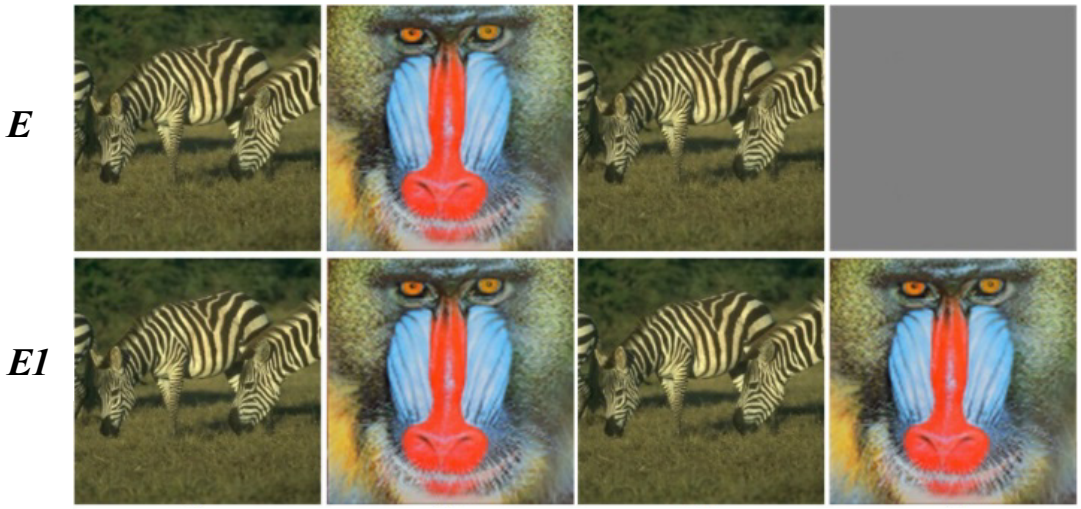
卓越的性能：本节将本章方法与基于 DWT-SVD-DCT[69]的传统图像水印技术进行了比较分析。受计算资源限制，本节仅探讨了在 de-raining 任务中嵌入 MDPI 水印的案例。实验结果总结如表 4.11 所示。传统图像水印技术由于需要精心设计水印嵌入与提取流程，其水印容量通常受到限制。相较而言，本章提出的基于深度学习的生成式模型水印算法通过神经网络的强大能力，能够便捷地嵌入更大容量的水印。此外，与传统技术相比，本章提出的方法在空间图像质量、频域隐蔽性和水印提取质量等方面均表现出显著优势。

表 4.11 与传统图像水印算法相比较的结果

方法	任务	嵌入水印	水印尺寸	PSNR	FP	BER
本章	de-raining	MDPI	128^2	54.59 dB	0.1622	0.0010
文献[69]	de-raining	MDPI	32^2	47.62 dB	0.1790	0.1291

4.3.5 过拟合问题分析

本研究提出的假设是，从总损失函数中去除干净损失函数可能引起水印提取网络的过拟合问题，这可能导致网络错误地从任何输入图像中提取出水印。为了对该假设进行验证，本节设计并执行了一项实验。具体而言，本节训练了两个水印提取网络：E（包含干净损失函数）和 $E1$（不包含干净损失函数），以观察干净损失函数的缺失是否会导致过拟合问题。



(a)含水印图像　(b)提取信息　(c)不含水印图像　(d)提取信息

图 4.7 过拟合实验结果

实验结果如图 4.7 所示。最上面一行显示的是 $\boldsymbol{E}$ 的结果，第二行显示的是 $\boldsymbol{E}1$ 的结果。图 4.7(a)是含水印图像，图 4.7(b)是从含水印图像中提取的结果，图 4.7(c)是清洁图像，图 4.7(d)是从清洁图像中提取的结果。可以发现，将不含水印的干净图像输入到网络 $\boldsymbol{E}1$ 时，$\boldsymbol{E}1$ 会错误地从这些干净图像中提取出水印。相比之下，网络 $\boldsymbol{E}$ 能够准确地从含水印图像中提取水印，同时从干净图像中提取出噪声。因此，干净损失的引入有效避免了水印提取网络的过拟合问题，这保证了本研究所提出的版权保护方法的有效性。

4.4 本章小结

本文的第三章通过优化水印在输出图像中的分布，有效缓解了水印添加过程中产生的高频伪影。然而，该方法并未解决由水印嵌入网络的固有结构所引起的高频伪影问题，仍存在局限性和改进空间。为了全面考虑水印嵌入网络的固有结构及水印添加过程所导致的高频伪影问题，本章提出了一种高保真的生成式模型水印技术。本章所提出的技术避免了使用水印嵌入网络，从而有效地解决了与深度网络固有结构相关的高频伪影问题。此外，本章所提出的技术引入了一个专门产生频率扰动的网络，该网络生成的频域扰动将作为水印被添加至原始图像的频

域中。通过将水印扰动限制在图像的低频部分，有效防止了水印嵌入过程对图像高频区域的影响。因此，本章的方法消除了生成式模型水印技术中的高频伪影，并大幅提高了水印系统的保真度。

但是必须认识到，本章提出的框架无法抵御所有现实世界中可能出现的攻击，因为预测和防范每一种可能的攻击是不现实的。例如，本章提出的方法在面对加性噪声攻击时表现出一定的脆弱性，这是因为噪声的添加会导致频域显著变化，从而破坏了水印提取所依赖的一致性。即便如此，通过消除高频伪影来有效隐藏水印，可以显著提升水印系统的保真度和隐蔽性，从而降低水印系统遭受攻击的可能性。因此，良好的隐蔽性能够显著降低水印被发现和攻击的概率，从而提高知识产权保护的安全性。

第五章 总结与展望

5.1 总结

身处于人工智能时代，以深度学习为核心的人工智能技术正在迅速发展，并已在多个领域获得了广泛的应用。深度学习模型的商业价值和重要性正在不断增加，而开发一个高性能的深度学习模型需要投入大量的资源。因此，保护深度学习模型的知识产权变得尤为重要，这一研究领域已成为学术界和产业界的关注焦点。本文研究面向生成式模型的模型水印技术。特别地，本文致力于解决生成式模型水印技术在水印图像的频域引入高频伪影的问题，这些高频伪影极大地降低了生成式模型水印的保真度。因此，为了消除生成式模型水印技术中的高频伪影并实现高保真的生成式模型水印技术，本文进行了以下三个方面的工作：

1) 概述了模型水印技术的研究背景、研究意义以及国内外研究现状。为了便于读者更好地理解，本文阐述了生成式模型水印的定义，并进一步介绍了该技术的评价指标及其面临的主要问题。此外，本文还介绍了高频伪影研究的相关背景知识。

2) 为了解决生成式模型水印中由水印嵌入过程所引起的高频伪影问题，本文提出了一种高保真的生成式模型水印框架。该框架采用了基于小波变换的频域分离层，能够有效地将图像按不同频率成分分解。通过对水印嵌入网络和水印提取网络进行联合训练以及优化联合损失函数，本框架成功地将水印嵌入到图像的低频区域，显著的减少了高频伪影。实验结果显示，该方法在不同任务中均取得了良好的效果。

3) 尽管上述研究有效地解决了水印嵌入过程中产生的高频伪影问题，但它未充分考虑由水印嵌入网络所引起的高频伪影。因此，本文进一步考虑了嵌入网络导致的高频伪影，并提出了一种高保真的生成式模型水印框架。本文提出的新框架不使用传统的水印嵌入网络，以避免由该结构产生的高频伪影。此外，本文

设计了一个频域扰动生成网络，该网络产生的扰动作为水印加入到图像的低频成分中，极大地降低了水印嵌入过程对图像高频区域的影响。实验结果表明，该方法在不同任务中均取得了良好的效果。相比之前的方法，本方法有效地消除了高频伪影，并显著提高了保真度和隐蔽性。

5.2 展望

本文旨在保护图像生成式网络的知识产权。本文深入分析了现有生成式模型水印技术，并重点揭示了这些技术在频域上所面临的高频伪影问题。为了解决高频伪影问题，本文提出了两种高保真的生成式模型水印技术，并取得了相关的学术成果。尽管这些研究成果在一定程度上推进了生成式模型水印技术的发展，但模型水印领域中仍存在许多未解决的问题。未来的研究可以从以下几个方向进行展开：

1) 针对大模型的模型水印技术：大型语言模型，例如 ChatGPT 等，由于其卓越的性能，已经获得了广泛的关注。这些模型主要功能是生成文本内容。当前的研究重点集中在如何保护这些大型语言模型的知识产权以及如何验证某段文本确实由特定的大型语言模型生成。这两个问题预计将成为未来研究的关键领域。

2) 数据集的知识产权保护：尽管当前深度学习的训练主要依赖开源数据集，但私有数据集同样值得关注。这些私有数据集通常具有重要价值，并可能包含敏感信息，例如医疗数据集，其泄漏可能引发严重的隐私安全问题。此外，一旦私有数据集泄露，任何人都能利用这些数据训练自己的深度学习模型。因此，为了解决私有数据集泄露的问题，开发一种能够追溯模型训练数据来源的水印技术显得尤为重要。

3) 结合模型可解释性的研究：通过交叉研究模型可解释性，研究人员不仅可以更合理地设计水印算法，还可能推导出模型可解释性的新理论。这种跨领域的相互借鉴是一个值得探索的研究方向。

参考文献

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

[2] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012: 25.

[3] ZHANG W, ZHAI M, HUANG Z, et al. Towards end-to-end speech recognition with deep multipath convolutional neural networks[C]//Proceedings of the Intelligent Robotics and Applications, August 8-11, 2019, Shenyang, China. Berlin: Springer, 2019: 332-341.

[4] 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.

[5] CHEN C, SEFF A, KORNHAUSER A, et al. Deepdriving: learning affordance for direct perception in autonomous driving[C]//Proceedings of the IEEE International Conference on Computer Vision, December 13-16, 2015, Santiago, Chile. Piscataway: IEEE, 2015: 2722-2730.

[6] RONNEBERGER O, FISCHER P, BROX T. U-net: convolutional networks for biomedical image segmentation[C]// Proceedings of the Medical Image Computing and Computer-Assisted Intervention, October 5-9, 2015, Munich, Germany. Berlin: Springer, 2015: 234-241.

[7] RAMESH A, DHARIWAL P, NICHOL A, et al. Hierarchical text-conditional image generation with clip latents[J]. Arxiv Preprint Arxiv:2204.06125, 2022.

[8] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF Conference

on Computer Vision and Pattern Recognition, June 19-20, 2022, New Orleans, America. Piscataway: IEEE, 2022: 10684-10695.

[9] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.

[10] LI Y, WANG H, BARNI M. A survey of deep neural network watermarking techniques[J]. Neurocomputing, 2021, 461: 171-193.

[11] 冯乐，朱仁杰，吴汉舟，等. 神经网络水印综述[J].应用科学学报，2021, 39(6): 881-892.

[12] 吴汉舟，张杰，李越，等. 人工智能模型水印研究进展[J].中国图象图形学报, 2023, 28(06):1792-1810.

[13] 张杰. 多重攻击下的深度模型水印方法研究[D]．合肥：中国科学技术大学，2022.

[14] UCHIDA Y, NAGAI Y, SAKAZAWA S, et al. Embedding watermarks into deep neural networks[C]//Proceedings of the ACM on International Conference on Multimedia Retrieval, June 6-9, 2017, Bucharest, Romania. New York: ACM, 2017: 269-277.

[15] DARVISH ROUHANI B, CHEN H, KOUSHANFAR F. Deepsigns: an end-to-end watermarking framework for ownership protection of deep neural networks[C]//Proceedings of the 24th International Conference on Architectural Support for Programming Languages and Operating Systems, April 13-17, 2019, Providence, America. New York: ACM, 2019: 485-497.

[16] FENG L, ZHANG X. Watermarking neural network with compensation mechanism[C]// Proceedings of the Knowledge Science, Engineering and Management, August 28-30, 2020, Hangzhou, China. Berlin: Springer, 2020: 363-375.

[17] WANG T, KERSCHBAUM F. Riga: covert and robust white-box watermarking of deep neural networks[C]//Proceedings of the Web Conference, April 19-23, 2021, Ljubljana, Slovenia. New York: ACM, 2021: 993-1004.

[18] CORTIÑAS-LORENZO B, PÉREZ-GONZÁLEZ F. Adam and the ants: on the influence of the optimization algorithm on the detectability of dnn watermarks[J]. Entropy, 2020, 22(12): 1379.

[19] LI Y, TONDI B, BARNI M. Spread-transform dither modulation watermarking of deep neural network[J]. Journal of Information Security and Applications, 2021, 63: 103004.

[20] CHEN H, ROHANI B D, KOUSHANFAR F. Deepmarks: a digital fingerprinting framework for deep neural networks[J]. Arxiv Preprint Arxiv:1804.03648, 2018.

[21] LOU X, GUO S, LI J, et al. Ownership verification of dnn architectures via hardware cache side channels[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(11): 8078-8093.

[22] ZHAO X, YAO Y, WU H, et al. Structural watermarking to deep neural networks via network channel pruning[C]// Proceedings of the IEEE International Workshop on Information Forensics and Security, December 7-10, 2021, Montpellier, France. Piscataway: IEEE, 2021: 1-6.

[23] FAN L, NG K W, CHAN C S. Rethinking deep neural network ownership verification: embedding passports to defeat ambiguity attacks[J]. Advances in Neural Information Processing Systems, 2019, 32.

[24] CHEN H, ROUHANI B D, KOUSHANFAR F. Specmark: a spectral watermarking framework for ip protection of speech recognition systems[C]//Proceedings of the Interspeech, October 25-29, 2020, Shanghai, China. Berlin: Springer, 2020: 2312-2316.

[25] ZHANG J, CHEN D, LIAO J, et al. Passport-aware normalization for deep model protection[J]. Advances in Neural Information Processing Systems, 2020, 33: 22619-22628.

[26] LIM J H, CHAN C S, NG K W, et al. Protect, show, attend and tell: empowering image captioning models with ownership protection[J]. Pattern Recognition, 2022, 122: 108285.

[27] ADI Y, BAUM C, CISSE M, ET AL. Turning your weakness into a strength: watermarking deep neural networks by backdooring[C]//Proceedings of the 27th USENIX Security Symposium, August 15-17, 2018, Baltimore, America. Berkeley，USENIX, 2018: 1615-1631.

[28] ZHANG J, GU Z, JANG J, ET AL. Protecting intellectual property of deep neural networks with watermarking[C]//Proceedings of the Asia Conference on Computer and Communications Security, June 4, 2018, Incheon, Korea. New York: ACM, 2018: 159-172.

[29] GUO J, POTKONJAK M. Watermarking deep neural networks for embedded systems[C]//Proceedings of the IEEE/ACM International Conference on Computer-Aided Design, November 5-8, 2018, San Diego, America. Piscataway: IEEE, 2018: 1-8.

[30] GUO J, POTKONJAK M. Evolutionary trigger set generation for dnn black-box watermarking[J]. Arxiv Preprint Arxiv: 1906.04411, 2019.

[31] LE MERRER E, PEREZ P, TRÉDAN G. Adversarial frontier stitching for remote neural network watermarking[J]. Neural Computing and Applications, 2020, 32: 9233-9244.

[32] CHEN H, ROUHANI B D, KOUSHANFAR F. Blackmarks: blackbox multibit watermarking for deep neural networks[J]. Arxiv Preprint Arxiv:1904.00344, 2019.

[33] SAKAZAWA S, MYODO E, TASAKA K, et al. Visual decoding of hidden watermark in trained deep neural network[C]//Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval, March 28-30, 2019, San Jose, America. Piscataway: IEEE, 2019: 371-374.

[34] SZYLLER S, ATLI B G, MARCHAL S, et al. Dawn: dynamic adversarial watermarking of neural networks[C]//Proceedings of the 29th ACM International Conference on Multimedia, October 20-24, 2021, Virtual. New York: ACM, 2021: 4417-4425.

[35] JIA H, CHOQUETTE-CHOO C A, CHANDRASEKARAN V, et al. Entangled watermarks as a defense against model extraction[C]//Proceedings of the 30th USENIX Security Symposium, August 11–13, 2021, Virtual. Berkeley, USENIX, 2021: 1937-1954.

[36] NAMBA R, SAKUMA J. Robust watermarking of neural network with exponential weighting[C]//Proceedings of the ACM Asia Conference on Computer and Communications Security, July 9-12, 2019, Auckland, New Zealand. New York: ACM, 2019: 228-240.

[37] LI H, WENGER E, SHAN S, et al. Piracy resistant watermarks for deep neural networks[J]. Arxiv Preprint Arxiv:1910.01226, 2019.

[38] LI F, WANG S. Knowledge-free black-box watermark and ownership proof for image classification neural networks[J]. Arxiv Preprint Arxiv:2204.04522, 2022.

[39] CAO X, JIA J, GONG N Z. Ipguard: protecting intellectual property of deep neural networks via fingerprinting the classification boundary[C]//Proceedings of the ACM Asia Conference on Computer and Communications Security, June 7-11,2021, Virtual. New York: ACM, 2021: 14-25.

[40] QUAN Y, TENG H, CHEN Y, et al. Watermarking deep neural networks in image processing[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(5): 1852-1865.

[41] ZHAO X, WU H, ZHANG X. Watermarking graph neural networks by random graphs[C]//Proceedings of the International Symposium on Digital Forensics and Security, June 28-29, 2021, Elazig, Turkey. Piscataway: IEEE, 2021: 1-6.

[42] TEKGUL B G A, XIA Y, MARCHAL S, et al. Waffle: watermarking in federated learning[C]// Proceedings of the International Symposium on Reliable Distributed Systems, September 20-23, 2021, Virtual. Piscataway: IEEE, 2021: 310-320.

[43] WU H, LIU G, YAO Y, et al. Watermarking neural networks with watermarked images[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(7): 2591-2601.

[44] ZHANG J, CHEN D, LIAO J, et al. Model watermarking for image processing networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence, February 7-12, 2020, New York, America. Menlo Park: AAAI, 2020: 12805-12812.

[45] ZHANG J, CHEN D, LIAO J, et al. Deep model intellectual property protection via deep watermarking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(8): 4005-4020.

[46] ZHANG L, LIU Y, SHAOTENG LIU, et al. Generative model watermarking based on human visual system[C]// Proceedings of the International Forum on Digital TV and Wireless Multimedia Communication, December 8-9, 2022, Virtual. Berlin: Springer, 2022: 136-149.

[47] LI Z, HU C, ZHANG Y, et al. How to prove your model belongs to you: a blind-watermark based framework to protect intellectual property of dnn[C]//Proceedings of the 35th Annual Computer Security Applications Conference, December 9-13, Puerto Rico, America. New York: ACM, 2019: 126-137.

[48] ODENA A, DUMOULIN V, OLAH C. Deconvolution and checkerboard artifacts [EB/OL]. (2016-10-17) Https://Distill.Pub/2016/Deconv-Checkerboard/.

[49] FRANK J, EISENHOFER T, SCHÖNHERR L, et al. Leveraging frequency analysis for deep fake image recognition[C]// Proceedings of the International Conference on Machine Learning, July 13-18, 2020, Virtual. Cambridge: JLMR, 2020: 3247-3258.

[50] ZENG Y, PARK W, MAO Z M, et al. Rethinking the backdoor attacks' triggers: a frequency perspective[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, October 10-17, 2021, Montreal, Canada. Piscataway: IEEE, 2021: 16473-16481.

[51] ZHANG C, BENZ P, KARJAUV A, et al. Udh: universal deep hiding for steganography, watermarking, and light field messaging[J]. Advances in Neural Information Processing Systems, 2020, 33: 10223-10234.

[52] FRIDRICH J. Digital image forensics[J]. IEEE Signal Processing Magazine, 2009, 26(2): 26-37.

[53] BURTON G J, MOORHEAD I R. Color and spatial structure in natural scenes[J]. Applied Optics, 1987, 26(1): 157-170.

[54] TOLHURST D J, TADMOR Y, CHAO T. Amplitude spectra of natural images[J]. Ophthalmic and Physiological Optics, 1992, 12(2): 229-232.

[55] WU H, LI C, LIU G, et al. Hiding data hiding[J]. Pattern Recognition Letters, 2023, 165: 122-127.

[56] ISOLA P, ZHU J Y, ZHOU T, et al. Image-to-image translation with conditional adversarial networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, Honolulu, America. Piscataway: IEEE, 2017: 1125-1134.

[57] LIANG W, HUANG W, LONG J, et al. Deep reinforcement learning for resource protection and real-time detection in iot environment[J]. IEEE Internet of Things Journal, 2020, 7(7): 6392-6401.

[58] LIU J, HAN J, FU K, et al. Application of qr code watermarking and encryption in the protection of data privacy of intelligent mouth opening trainer[J]. IEEE Internet of Things Journal, 2023, 10(12): 10510-10518.

[59] ZHANG C, BENZ P, KARJAUV A, et al. Universal adversarial perturbations through the lens of deep steganography: towards a fourier perspective[C]//Proceedings of The AAAI Conference on Artificial Intelligence, May 19-21, 2021, Virtual. Menlo Park: AAAI, 2021, 35(4): 3296-3304.

[60] BRANWEN G, GOKASLAN A. Danbooru2019: a large-scale crowdsourced and tagged anime illustration dataset [EB/OL]. (2022-11-1) Https://Gwern.Net/Danbooru2021#Danbooru2019.

[61] YANG W, TAN R T, FENG J, et al. Deep joint rain detection and removal from a single image[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, Honolulu, America. Piscataway: IEEE, 2017: 1357-1366.

[62] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of the IEEE International Conference on Computer Vision, October 22-29, Venice, Italy. Piscataway: IEEE, 2017: 2223-2232.

[63] DAUBECHIES I. Orthonormal bases of compactly supported wavelets[J]. Communications on Pure and Applied Mathematics, 1988, 41(7): 909-996.

[64] JOHNSON J, ALAHI A, FEI-FEI L. Perceptual losses for real-time style transfer and super-resolution[C]//Proceedings of the European Conference on Computer Vision, October 11-14, 2016, Amsterdam, Netherlands. Berlin: Springer, 2016: 694-711.

[65] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[66] WANG Z, SIMONCELLI E P, BOVIK A C. Multiscale structural similarity for image quality assessment[C]//Proceedings of the 37th Asilomar Conference on

Signals, Systems & Computers, November 09-12, 2003, Pacific Grove, America. Piscataway: IEEE, 2003, 2: 1398-1402.

[67] BALUJA S. Hiding images in plain sight: deep steganography[J]. Advances in Neural Information Processing Systems, 2017, 30.

[68] ZHANG R, ISOLA P, EFROS A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, America. Piscataway: IEEE, 2018: 586-595.

[69] HE Y, HU Y. A proposed digital image watermarking based on dwt-dct-svd[C]//Proceedings of the 2th IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference, May 25-27, 2018, Xi'an, China. Piscataway: IEEE, 2018: 1214-1218.

[70] ZHANG L, LIU Y, ZHANG X, et al. Generative model watermarking suppressing high-frequency artifacts[J]. Arxiv Preprint Arxiv:2305.12391, 2023.

攻读硕士学位期间取得的研究成果

一、 论文

（*共同一作）

[1] Liu Y*, **Zhang L***, Wu H, Wang Z, Zhang X. Reducing high-frequency artifacts for generative model watermarking via wavelet transform[J]. IEEE Internet of Things Journal, 2024. (CCF C，SCI 一区，IF:10.6)

[2] **Zhang L**, Zhang X, Wu H. High-frequency artifacts-resistant image watermarking applicable to image processing models[J]. Applied Sciences. 2024. (SCI 三区，IF:2.7)

[3] **Zhang L***, Liu Y*, Liu S, Yang T, Wang Y, Zhang X, Wu H. Generative model watermarking based on human visual system[C]//Proceedings of the International Forum on Digital TV and Wireless Multimedia Communication, 2022. (EI)

[4] **Zhang L***, Liu Y*, Zhang X, Wu H. Suppressing high-frequency artifacts for generative model watermarking by anti-aliasing[C]//Proceedings of the ACM Workshop on Information Hiding and Multimedia Security, 2024. (CCF C，信息隐藏与多媒体安全领域顶级国际学术会议)

致　谢

随着毕业论文的完成，我的研究生生涯也即将告一段落。在此，我要向所有在我研究生期间给予我帮助和支持的人表示衷心的感谢。

首先，我要感谢我的导师张新鹏教授。在我研究生期间，张老师一直给予我悉心的指导和无私的帮助。张老师严谨的科研态度和深厚的学术造诣深深地影响了我，使我受益匪浅。另外，我还要感谢吴汉舟老师，从论文选题到实验设计，再到论文撰写，每一个环节吴老师都深度参与并悉心指导，给予我许多宝贵的意见和建议。在此我向张老师和吴老师表示深深的敬意和感谢。

其次，我要感谢实验室的师兄师姐和同门。他们在科研过程中给予了我很多帮助，无论是科研想法的交流，还是实践技术的指导，都让我感受到了实验室大家庭的温暖。与他们的交流与合作，使我在学术上不断进步，更加明确自己的前进的方向。

最后，我要特别感谢我的父母和女朋友。他们一直是我最坚实的后盾，无论我遇到什么困难和挫折，都始终给予我最大的支持和鼓励。是他们的无私奉献和默默付出，让我能够专心投入到学术研究中，不断取得新成果。祝愿他们身体健康，万事如意。

最后，感谢抽出宝贵时间评阅论文的专家老师们，向你们表示我诚挚的谢意！

作者署名：张力

完成地点：上海大学

2024 年 5 月 10