中图分类号：　　单位代号：10280
密　　级：　　学　　号：23721141

上海大学

硕士学位论文

SHANGHAI UNIVERSITY
MASTER'S DISSERTATION

题目	面向 KAN 模型版权保护的鲁棒水印技术研究

作　　者：赵一飞
学科专业：电路与系统
导　　师：吴汉舟
完成日期：2026 年 5 月

姓　　名：赵一飞　　学号：23721141

论文题目：面向 KAN 模型版权保护的鲁棒水印技术研究

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：赵一飞　　学号：23721141

论文题目：面向 KAN 模型版权保护的鲁棒水印技术研究

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

面向 KAN 模型版权保护的鲁棒水印技术研究

作　　者：赵一飞

学科专业：电路与系统

导　　师：吴汉舟

上海大学通信与信息工程学院

2026 年 5 月

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

Research on Robust Watermarking Techniques for Copyright Protection of KAN models

Candidate： **Yifei Zhao**

Major： **Circuits and Systems**

Supervisor： **Hanzhou Wu**

School of Communication and Information Engineering,

Shanghai University

May, 2026

摘　要

随着人工智能技术的快速发展，如何保护人工智能模型的知识产权成为学术界和工业界重点关注的课题。现有研究利用数字水印技术保护卷积神经网络、循环神经网络等主流人工智能模型的知识产权，能够较好平衡模型在原始任务和水印任务上的计算性能。作为一种新型架构，KAN 模型将可学习激活函数置于网络的“边”上，而非将固定的激活函数置于“节点”处，这一设计提升了其拟合复杂函数的能力，得到了越来越多的关注。然而，直接将现有方法用于 KAN 模型会导致水印的隐蔽性较差，且鲁棒性存在不足。在此背景下，本文开展了适用于 KAN 模型的鲁棒水印技术研究，取得的主要成果如下：

（1）KAN 模型由于其激活函数善于拟合连续平滑的函数，训练时会优先拟合样本量大、损失下降明显的原始任务，导致水印任务的学习能力存在不足。同时，KAN 模型的可调参数分布在激活函数上，参数冗余空间小于传统模型，进一步增大了水印嵌入难度。针对这一问题，本文利用频域扰动和交替训练实现黑盒 KAN 模型水印。该方案通过在频域添加扰动，使触发信号影响样本的整个空域，从而让模型更好地学习触发模式。同时，为了增强模型在训练时学习水印任务的能力，该方案采用交替训练的方式优化模型参数，并引入模型参数扰动损失以模拟攻击引发的参数变化。实验结果表明，该方案在保持模型原始任务性能的同时，有效提升了水印的检测性能，并可抵抗常见的模型水印攻击。

（2）上述方案缺乏对激活函数形变幅度施加有效约束，导致水印模型与干净模型对应边上的激活函数存在明显形态差异，削弱了水印的隐蔽性。针对这一问题，本文提出一种基于激活扰动的 KAN 模型水印方案。该方案仅对原始任务训练后的 KAN 模型第一层激活函数施加微小扰动来完成水印嵌入，并约束其形变幅度以保持形态一致性。同时，为了增强水印验证网络的抗攻击能力，该方案引入参数对抗训练，通过在模型激活函数中施加随机噪声来模拟攻击扰动，使水印验证网络在模型遭受攻击后仍能成功提取水印信息。实验结果表明，该方案可实现水印的高隐蔽性嵌入，并能在常见模型水印攻击后准确提取水印。

关键词：模型水印；KAN 模型；版权保护；鲁棒性；隐蔽性

ABSTRACT

With the rapid development of artificial intelligence technology, how to protect the intellectual property rights of artificial intelligence models has become a key concern in both academia and industry. The existing research utilizes digital watermarking techniques to protect the intellectual property of mainstream artificial intelligence models such as convolutional neural networks and recurrent neural networks, achieving a good balance between model performance on the original task and the watermarking task. As a novel architecture, the KAN model places learnable activation functions on the "edges" of the network instead of fixed activation functions at the "nodes", which enhances its ability to fit complex functions and has received increasing attention. However, directly applying existing methods to the KAN model will result in poor watermark concealment and insufficient robustness. In this context, this dissertation conducts research on robust watermarking techniques applicable to KAN models, and the main contributions are as follows:

(1) Due to the activation function of KAN models being proficient at fitting continuous smooth functions, they tend to prioritize fitting the original task with large sample sizes and significant loss reduction during the training process, resulting in insufficient learning ability for the watermarking task. Meanwhile, since the adjustable parameters of KAN models are distributed across activation functions, the parameter redundancy space of KAN models is smaller compared to traditional models, thereby further increasing the difficulty of watermark embedding. To address this issue, this dissertation utilizes frequency domain perturbation and alternating training to achieve black-box KAN model watermarking. This method adds perturbations in the frequency domain to affect the entire spatial domain of samples, thereby enabling the model to better learn trigger patterns. Simultaneously, to enhance the ability of models to learn the watermarking task during training, this method employs an alternating training strategy to optimize model parameters and introduces model parameter perturbation loss to simulate parameter changes caused by attacks. Experimental results demonstrate that the proposed method maintains the performance of models on the original task while effectively improving watermark detection performance and resisting common model watermarking attacks.

(2) The aforementioned method lacks effective constraints on the deformation magnitude of activation functions, resulting in noticeable morphological differences in the activation functions on corresponding edges between the watermarked model and the clean model, thereby undermining the stealthiness of the watermark. To address this issue, this dissertation proposes a KAN model watermarking method based on activation perturbation. This method embeds watermarks by applying small perturbations solely to the first-layer activation functions after the model completes training on the original task, while constraining the deformation magnitude of activation functions to preserve morphological consistency. Simultaneously, to enhance the attack resistance capability of the watermark verification network, this method introduces parameter adversarial training, where random noise is applied to the model's activation functions to simulate attack perturbations, enabling that the watermark verification network can successfully extract watermark information even after the model has been attacked. Experimental results demonstrate that the proposed method achieves highly imperceptible watermark embedding and enables accurate watermark extraction after common model watermarking attacks.

Keywords: Model Watermarking; KAN Models; Copyright Protection; Robustness; Invisibility

目　录

第一章　绪论

1.1 研究背景与意义

近年来，随着深度学习(Deep Learning, DL)[1]技术的迅猛发展，基于深度神经网络(Deep Neural Network, DNN)的人工智能(Artificial Intelligence, AI)技术实现了从基础研究到工程应用的重大跨越，已经被广泛应用于日常生活与工业生产领域，有力推动了社会进步与经济发展。目前，深度神经网络模型已经在计算机视觉[2]、自然语言处理[3]、自动驾驶[4]等诸多领域取得卓越表现，为人们生活带来便利的同时，持续赋能社会发展，故而高性能神经网络模型的研发已成为学术界与工业界共同关注的研究重点。然而，构建高性能的神经网络模型通常需要大规模高质量的标注数据、精巧的架构设计、专业化的参数调优以及庞大的计算资源与时间成本，这使得训练好的高性能模型成为极具价值的知识资产。基于此，模型的销售或租赁已成为一种可行的商业模式[5]，但这种商业服务模式也面临模型被非法转售的风险，可能对版权拥有者造成重大经济损失。因此，保护模型知识产权已成为亟待解决的关键问题。

实际上，随着神经网络在实际应用中的日益普及，其面临的安全威胁也愈发严峻，深度模型的知识产权保护已引起全球广泛关注。国际层面，欧盟《人工智能法案》于 2024 年 8 月 1 日正式生效，成为全球范围内具有标志性意义的人工智能综合性监管法规。2025 年 7 月 26 日，世界人工智能大会暨人工智能全球治理高级别会议在上海开幕，会议发布了《人工智能全球治理行动计划》，提出 13 项切实可行的具体行动，推动形成人工智能全球治理框架和规则。国内层面，我国人工智能安全治理体系在 2024-2025 年持续完善：2024 年 3 月，《生成式人工智能服务安全基本要求》首次系统性明确语料安全、模型安全、安全措施、风险词库建设及安全评估等全流程技术要求。2025 年 3 月，国家互联网信息办公室等四部门联合发布《人工智能生成合成内容标识办法》，旨在规范人工智能生成合成内容标识。2025 年 9 月，《人工智能安全治理框架》2.0 版正式发布，细化

三类安全风险，强化全生命周期技术治理。此外，2025 年 3 月，全国两会期间，政府工作报告多次提到了人工智能，明确提出持续推进“人工智能+”行动，国务院《关于深入实施“人工智能+”行动的意见》专设“提升安全能力水平”的条款，要求推动模型算法、数据资源、基础设施、应用系统等安全能力建设，加快形成动态敏捷、多元协同的人工智能治理格局。随着一系列政策法规的出台，标志着模型知识产权保护已从学术议题上升为国家战略层面的重要任务。

早期的版权保护研究主要以多媒体数据格式的数字资产为对象[6-8]，其中数字水印[9,10]技术是保护数字资产安全的主流方法之一。该技术通过将特定信息作为水印嵌入图像[11]、音频[12]等多媒体产品中，版权拥有者可在可疑数字产品中提取水印信息以证明侵权行为。然而，与以多媒体内容为主要载体的数字资产保护不同，深度学习模型作为一种新兴的知识载体，无法直接将传统数字水印技术照搬到其知识产权保护中。为解决这一问题，研究者们将数字水印思想拓展至深度学习领域，发展出适用于神经网络模型知识产权保护的模型水印技术[13]，以满足深度学习模型的可追溯性验证需求。根据嵌入和提取水印时对模型访问权限的不同，神经网络水印技术大致可以分为白盒水印、黑盒水印和无盒水印三种。

然而，将现有模型水印方案直接应用于 KAN 模型时，仍然会面临着诸多挑战。KAN 模型的激活函数因其擅长拟合连续平滑函数，在训练过程中会优先拟合样本量大、损失下降显著的原始任务，从而导致 KAN 模型对水印任务的学习能力相对薄弱。同时，与传统神经网络不同，KAN 模型的参数分布于各边的可学习激活函数上，水印嵌入过程直接体现为激活函数形态的改变，而现有方案在训练过程中主要以水印验证准确率和原始任务精度为目标，缺乏对激活函数形变幅度施加有效的约束，致使水印模型与干净模型在对应边上的激活函数呈现明显形态差异，严重削弱了水印隐蔽性。因此，设计面向 KAN 模型的鲁棒性和隐蔽性水印方案是本文研究的重点。

1.2 国内外研究概况

模型水印技术作为目前主流的深度模型知识产权保护方法，被国内外众多研

究人员广泛研究[14,15]。根据模型所有权验证阶段所需的条件不同，模型水印方法大致可以分为白盒水印、黑盒水印和无盒水印三类。白盒水印通常将水印信息直接嵌入到目标模型的参数或结构中。在验证阶段，模型所有者需要全面访问目标模型的内部细节，包括网络架构和模型参数，以便准确提取水印信息，从而验证目标模型的所有权。与白盒水印技术相比，黑盒水印技术无需模型所有者访问目标模型的内部细节，仅通过与目标模型的交互即可实现所有权验证。相关研究表明，即使训练数据被错误标注，深度学习模型仍然能够很好地拟合这些异常关系[16]。黑盒水印技术利用模型的这种冗余特性，记忆一些额外的特殊输入与特定输出间的关系，并将其作为模型所有者的版权标记。因此，在水印嵌入阶段，模型所有者使用正常训练集和触发集对目标模型进行训练，最终获得一个带水印的模型。在验证阶段，模型所有者通过获取目标模型在触发集上的预测结果，并与预先指定的标签进行一致性分析，从而验证模型的所有权。与白盒水印和黑盒水印不同，无盒水印则是通过将模型输出与不可见水印相结合，使得受保护模型的输出携带水印信息。在模型所有权验证阶段，模型所有者通过提取输出中的水印信息来验证模型的所有权。

1.2.1 白盒水印

由于白盒水印通常是将水印信息直接嵌入到目标模型的参数或结构中，因而在验证阶段，模型所有者需要全面访问目标模型的内部细节，包括网络架构和模型参数，以便准确提取水印信息，从而验证目标模型的所有权。

Uchida 等人[17]首次提出了基于白盒方法的深度神经网络水印方案，为模型知识产权的主动保护开辟出了新的道路。该方法选择深度神经网络模型的权重参数作为水印信息的载体，在原始任务的损失函数中引入嵌入正则化项，通过联合优化实现水印的隐式嵌入。具体而言，首先生成一个密钥矩阵作为嵌入参数，然后在训练过程中约束模型权重与密钥矩阵的乘积经激活函数映射后的输出序列与待嵌入的二进制水印序列一致，从而实现水印嵌入；同时通过调整正则约束项的权重系数，确保水印嵌入不会显著损害模型在原始任务上的性能。在水印提取阶段，模型所有者利用密钥矩阵与模型权重进行计算即可恢复水印信息，从而验

证模型所有权。为了增强水印信息与模型输入之间的映射关系，避免嵌入的水印仅依赖模型静态权重，Rouhani 等人[18]在 Uchida 等人[17]方案的基础上，将特定的水印字符串嵌入到深度神经网络各层激活图的概率密度函数中。这种设计使水印信息与输入数据特征建立动态关联，在提升水印灵活性的同时，进一步增强了模型水印的鲁棒性与安全性。同样是选择模型内部参数作为嵌入位置，不同于 Uchida 等人[17]的随机权重选取，Liu 等人[19]提出一种称为“贪婪残差”的权重选择方案，通过贪婪地选取少量且重要的关键参数，并将水印嵌入到这些参数构造的残差值中，使嵌入的水印展现出较强的鲁棒性。

然而，上述白盒水印方案在目标模型嵌入水印后均会致使水印模型参数的权值分布与原始模型产生偏离，从而容易遭受基于统计分析的攻击检测。Wang 等人[20]正是利用此缺陷，通过分析权重分布的方差特征来探知水印长度，从而实施精准的覆盖攻击。为应对此威胁，Wang 等人[21]提出了 RIGA 白盒水印框架，将水印模型的训练过程作为生成器以及水印检测作为鉴别器进行对抗训练来约束水印模型的参数分布与原始模型保持一致，从而提升水印的隐蔽性。Kuribayashi 等人[22]则是摒弃了传统的损失函数嵌入方式，将量化索引调制技术引入深度神经网络模型水印领域，通过直接修改权重初始值以及在训练过程中逐周期校正的方式实现水印嵌入，从而严格控制水印嵌入引起的模型权重变化量，进而也提升了水印的隐蔽性。

除深度模型权重参数外，模型结构信息亦可作为水印信息嵌入的载体，并且此类基于结构的水印方案对参数修改攻击具备天然的鲁棒性。Zhao 等人[23]提出了一种使用通道修剪的结构化水印方法，将水印嵌入到主机架构中，而不是直接修改模型的参数，从而规避了常见的基于参数的攻击，并且显示出良好的应用前景。同样，Lou 等人[24]利用神经架构搜索来获得嵌入水印的模型结构，也表现出对基于参数修改的攻击的高鲁棒性。此外，Fan 等人[25]提出了一种基于数字护照的深度神经网络模型保护方法，通过在卷积层后添加“护照层”来解决传统模型水印无法抵抗混淆攻击的问题。该方法将护照层的缩放因子和偏置项设计为由卷积权重与护照信息共同决定，从而建立了模型性能与护照信息的强耦合关系。具体而言，当且仅当提供正确的护照信息时，模型才能维持正常的推理性能；一旦

护照被篡改或伪造，模型分类准确率将急剧下降。这一机制使得攻击者无法通过逆向工程来伪造有效护照，从而有效抵御混淆攻击。

综上所述，经过多年的发展，白盒水印正逐步从初步的概念探索走向更为系统的应用研究，展现出在模型知识产权保护方面的巨大潜力。从早期基于权重参数嵌入的隐式正则化方案，到利用激活图分布、残差权重选择等增强策略，再到结合对抗训练、量化索引调制技术来提升水印的隐蔽性，以及基于模型结构信息和数字护照等新型嵌入范式，白盒水印技术已经形成了多层次、多维度的发展格局。这些技术通过在模型内部嵌入隐蔽的标识信息，在保障模型推理性能不受显著影响的前提下，实现模型来源追溯与权属认证，进而保障深度神经网络模型的安全部署和知识产权。

1.2.2 黑盒水印

白盒水印要求模型所有者在验证过程中必须掌握可疑模型的内部细节，例如网络结构与权重参数等核心信息，以此完成水印提取及所有权确认。然而，这种对完全访问权限的硬性要求极大地束缚了其实际应用。为突破这一局限，研究者相继提出多种黑盒验证范式。该类方法的核心思想在于利用深度神经网络的冗余拟合能力，即便训练数据中存在部分标注错误，模型仍能有效学习这些异常映射关系[16]。黑盒水印正是基于这一特性，通过构造特殊的输入输出关联作为权属标识。在技术实现上，模型所有者首先选取或生成一组触发样本，并为其指定特定的预测标签；随后将触发样本与正常训练数据共同用于模型优化，使模型在保持原有任务性能的同时，隐式记忆这些特殊映射。验证时，只需向可疑模型输入触发样本并比对其输出与预设标签的一致性，即可完成所有权判定。

Adi 等人[26]巧妙地将后门攻击的思想转化为模型版权保护的技术路径，提出了一种简单有效的黑盒水印方案。该方法的核心在于利用深度神经网络的过度参数化特性，将水印嵌入过程设计为一种后门机制。具体而言，首先随机选取一组与目标任务无关的抽象图像，并为每个抽象图像分配预定义的目标标签以构建触发集；随后将触发集混合到原始训练集中，通过联合优化原始任务损失与后门触发损失完成模型训练。由此得到的模型在常规输入上维持原有性能表现，而在特

定触发输入上则输出预设的目标标签。在验证阶段，模型所有者仅需通过黑盒查询接口远程访问目标模型并输入特定的抽象触发图像，若模型输出与预设标签一致，即可证明模型的所有权。Zhang 等人[27]进一步探究了不同类型的触发后门的应用，包括无关图像、随机噪声与可见标记。这些方法有效提升了水印的隐蔽性与通用性，但也带来了新的挑战，例如无关图像后门较易被检测，而采用随机噪声则难以有效表征模型所有者的身份。Guo 等人[28]则提出了一种改进方法，采用以用户信息引导的噪声来作为触发样本，将不可感知的噪声与模型的所有者相关联，从而实现明确的模型所有权归属。虽然该方法能够提升水印的安全性与可解释性，但是也对触发样本的设计提出了更高的要求。随着数字水印技术的不断发展，研究人员逐渐倾向于在频域中嵌入水印信息，而不是仅仅局限于空域。尽管在空域中嵌入水印信息的方法由于其实现简单以及计算效率高而具有一定的优势，但其在鲁棒性方面的性能往往不尽如人意。相比之下，在频域中嵌入水印信息通常具备更优异的鲁棒性，从而能够更好地满足实际应用中对水印技术的严格要求。Liu 等人[29]选择在频域内进行水印信息的嵌入，借助傅里叶扰动分析探究了输入样本中不同频率分量对模型任务性能的影响，并通过 K 均值聚类算法确定了适用于触发样本制作的水印嵌入频率。实验结果表明，该方法在维持模型原始任务性能的前提下，实现了更优的水印效果。Mo 等人[30]也提出了一种在频域内嵌入身份水印信息的方法，该方法结合了离散余弦变换和奇异值分解，并引入了对抗训练策略来增强模型区分正确和不正确身份水印信息的能力，同样实现了较好的水印效果。

借助对模型自身特性的挖掘与利用，对抗样本方法也被应用于模型水印的构建当中，Merrer 等人[31]以原始模型决策边界附近的对抗样本为基础，通过微调模型决策边界来嵌入水印信息，使原本被错误分类的真实对抗样本能够被模型正确分类，从而实现模型的所有权认证。为了进一步提高水印的鲁棒性，Jia 等人[32]提出了一种纠缠水印嵌入策略。该方法通过引入软最近邻损失约束目标模型，使水印任务与正常任务共享相同的特征表示，从而将后门特征紧密耦合在正常特征表达中。当攻击者试图窃取模型时，其在学习原始特征的过程中也必然同时学习到后门特征，从而显著增强了模型抵御窃取攻击的能力。Charette 等人[33]则在模

型输出中注入诸如余弦分布这类特定的信号分布，让攻击者窃取得到的模型也会同步学习到该分布特征。此外，Li 等人[34]采用无目标后门攻击的方式实现水印嵌入，弱化了特定触发模式与目标标签之间的关联。Xu 等人[35]则是摒弃了为触发图像分配异常标签的传统方式，而是选用非分类的随机图像作为触发图像，并设计特定规则的序列号作为水印信息，通过非相关多任务学习将触发图像与序列号的映射关系嵌入目标模型，在规避异常标签带来的安全隐患的同时，有效提升了水印的嵌入容量。

黑盒水印是一种无需访问模型内部参数与结构，仅需通过模型输入输出行为即可完成模型所有权验证的技术。这一特性突破了传统白盒水印必须依赖模型内部访问权限的应用限制，使其具备更强的实用性与更广的适配性。随着该技术日趋成熟，其应用范围也从最初的图像分类模型，逐步拓展至图像处理网络[36]、图神经网络[37]等各类深度学习架构，并进一步覆盖图像处理[38]、图像生成[39]、语音识别[40]、自然语言处理[41]等多样化实际任务场景，为不同架构、不同场景下的深度学习模型版权认证提供了全面且有效的技术支撑。

1.2.3 无盒水印

无盒水印是近年来新兴的模型知识产权保护范式，其核心特征在于验证过程中既无需获取模型内部参数，也无需与目标模型进行任何交互查询。该类方法主要面向生成式模型的版权保护场景，通过向模型输出内容（如图像、文本、音频等）中嵌入不可感知的标识信息，实现权属追溯与侵权认定。

Wu 等人[42]提出了一种面向图像生成模型的无盒水印框架。该框架联合训练深度神经网络与水印提取网络，通过优化组合损失函数，使深度神经网络在完成原始任务的同时将不可感知的水印嵌入输出图像，并确保水印提取网络仅在提供正确密钥时方可从含水印图像中恢复预设水印，否则输出噪声，从而实现了图像内容与模型产权的双重保护。为增强模型抵御窃取攻击的能力，Zhang 等人[43]提出一种基于输出分布水印的模型保护方案，其核心思想是将预定义的版权信息以不可见形式嵌入图像处理模型输出的特征分布而非直接修改输出图像的像素内容，从而使攻击者在利用输入输出对训练代理模型时难以轻易剥离或破坏隐藏的

版权标识，并且能够通过相应的水印提取网络从可疑模型的输出中准确恢复原始水印信息，进而实现对模型知识产权的有效溯源。为了进一步应对攻击者在模型窃取时对数据进行预处理的情况，Zhang 等人[44]提出了一种基于结构一致性的模型水印框架。该方法将水印信息与图像中具有语义不变性的物理结构进行对齐嵌入，使得水印在经历旋转、裁剪等数据增强操作后仍能保持一致性，从而有效提升了对抗数据预处理攻击的鲁棒性。然而该方法会在标记图像的高频频谱中引入高频伪影，显著降低水印的隐蔽性与水印系统的安全性。针对这一问题，Zhang 等人[45]摒弃了传统水印技术中常用作水印嵌入网络的卷积神经网络结构，设计了一种频率扰动生成网络，将水印信息转化为低频扰动并嵌入载体图像经傅里叶变换后的低频频谱区域，避免对图像高频分量造成修改，从而有效抑制了高频伪影的产生，大幅提升了水印的隐蔽性。Liu 等人[46]则是借助离散小波变换，在水印嵌入网络后增设小波频率分离层，将其生成的图像分解为不同频率分量，并通过联合训练与损失优化，将水印信息约束嵌入至图像低频区域，既有效抑制了高频伪影的产生，也让水印在频域中的分布更贴合自然图像的特征。

相较于需获取模型内部细节的白盒水印和需直接交互模型的黑盒水印，无盒水印仅需要通过分析模型的最终输出即可完成模型版权的验证。它通过在模型生成的图像、文本等内容中嵌入不可感知的水印标识信息，使验证者能够仅依据可疑模型的输出便能够实现权属追溯。本质上，无盒水印可看作黑盒水印在生成式模型上的特殊情况，其提取端获取的信息更少，却为生成式内容版权保护提供了切实可行的技术路线。

1.3 研究内容和结构安排

1.3.1 研究内容

本论文围绕 KAN 模型的知识产权保护问题展开研究。在深度神经网络知识产权保护日益受到关注的背景下，KAN 模型作为一种具备替代多层感知机潜力的新兴架构，其知识产权保护问题同样不容忽视。与多层感知机不同，KAN 模型采用了独特的架构设计：将可学习的激活函数置于网络"边"而非"节点"上，这

一创新设计显著提升了模型拟合复杂函数的能力，但也赋予了其与传统网络截然不同的鲜明特性，对现有模型保护方案提出了新的挑战。

当前，研究者们已将数字水印技术拓展至模型版权保护领域，在模型盗版检测、所有权认证及篡改追踪等方面取得了阶段性进展。然而，将现有方案直接应用于 KAN 模型时，仍面临诸多挑战，如面对水印去除攻击时的鲁棒性不足、水印嵌入的隐蔽性欠缺以及嵌入过程计算成本高昂等问题。针对上述问题，本文提出了两种面向 KAN 模型的水印方案。具体工作如下：

由于 KAN 模型的激活函数善于拟合连续平滑函数，其在训练过程中会优先拟合样本量大、损失下降明显的原始任务，从而导致水印任务的学习能力存在不足。同时，由于 KAN 模型的可调参数分布于激活函数，导致 KAN 模型的参数冗余空间较传统模型更小，使嵌入水印的难度增大。针对这一问题，本文提出了一种基于频域扰动和交替训练的鲁棒黑盒 KAN 模型水印方案。该方案首先将水印信息嵌入图像频域以构建触发集；继而采用交替训练策略，轮流使用原始训练集与触发集进行模型训练，使模型在学习原始任务的同时有效习得水印任务；此外，在训练过程中引入模型参数扰动损失以模拟水印去除攻击场景，从而确保水印模型在遭受常见去除攻击后仍能可靠完成水印验证任务。实验结果表明，该方案在保持模型原有任务性能的同时，有效增强了水印的鲁棒性。

上述方案在训练过程中主要以水印验证准确率和原始任务精度为目标，缺乏对激活函数形变幅度施加有效的约束，致使水印模型与干净模型对应边上的激活函数存在显著形态差异，削弱了水印的隐蔽性。针对这一问题，本文提出了一种基于激活扰动的 KAN 模型水印方案。该方案仅对原始任务训练后的 KAN 模型第一层激活函数施加轻微扰动以嵌入水印信息，有效提升了水印在激活函数层面的形态不可感知性，同时避免了对模型进行完整的重新训练，有效降低了计算成本。为了增强水印验证网络的抗攻击能力，该方案引入了参数对抗训练，通过对 KAN 模型激活函数施加特定程度的随机噪声以模拟水印去除攻击场景，使水印验证网络在模型遭受攻击后仍能成功提取水印信息。实验结果表明，该方案在保持嵌入便利性的同时，实现了水印的高隐蔽性嵌入，并确保水印验证网络在常见的模型水印去除攻击后仍然能够准确提取水印，从而使水印在隐蔽性与鲁棒性之

间取得了良好的平衡。

1.3.2 结构安排

本论文总共分为五个章节，各章内容安排如下：

第一章阐述了 KAN 模型水印的研究背景和意义，并着重介绍了模型水印技术的国内外研究现状。

第二章阐述了面向图像分类任务的 KAN 模型水印相关技术基础。首先介绍 KAN 模型的理论根基与网络结构特性，梳理其变体在激活函数上的创新，随后详解三种频域嵌入方法的数学原理及水印嵌入机制，最后梳理模型水印的分类与定义，分析常见水印攻击机理，明确三大评价指标内涵，为后续 KAN 模型鲁棒水印方案构建提供理论支持。

第三章为基于频域扰动和交替训练的 KAN 模型水印方法，首先介绍了该水印方案的整体流程以及 KAN 模型的训练过程。然后阐述了交替训练策略的设计动机，并引入模型参数扰动损失来模拟剪枝和微调等水印去除攻击，从而提升了水印的鲁棒性。最后，对该水印框架的算法性能进行了实验评估，包括保真度分析、鲁棒性分析、对比实验、消融实验和计算成本分析。

第四章为基于激活扰动的 KAN 模型水印方法，首先介绍了所提出方法的整体框架，然后介绍了水印的嵌入与提取的流程以及在训练过程中所使用的损失函数，最后对该水印框架的算法性能进行了实验评估，包括保真度分析、不可见性分析，面对常见水印去除攻击的鲁棒性分析和对比实验。

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 KAN 模型水印的研究背景和研究意义，同时介绍了国内外模型水印技术的研究现状，并在最后介绍了本文的主要研究内容和结构安排。

第二章　相关技术基础

2.1 KAN 模型

2.1.1 KAN 模型的基本概念

KAN 模型的提出源自于 Kolmogorov-Arnold 表示定理(Kolmogorov-Arnold Representation Theorem, KART)[47-49]的理论支撑，该定理指出，任何在有界域上定义的连续多元函数f都可以被表示为通过加法组合的连续单变量函数的有限复合。对于一组变量$x = x_1, x_2, \ldots, x_n$，其中$n$是变量的数量，连续多元函数$f(x)$可以表示为公式(2.1)：

$$f(x) = f(x_1, x_2, \ldots, x_n) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^{n} \phi_{q,p}(x_p) \right) \tag{2.1}$$

其中 $\phi_{q,p} : [0,1] \to \mathbb{R}$ 并且 $\Phi_q : \mathbb{R} \to \mathbb{R}$ 。

一个包含L层的多层感知机(Multi-layer Perceptron, MLP)[50-52]可以被描述为变换矩阵W和激活函数σ的相互作用，其数学表达式为公式(2.2)：

$$MLP(x) = (W_{L-1} \circ \sigma \circ W_{L-2} \circ \sigma \circ \cdots \circ W_1 \circ \sigma \circ W_0)x \tag{2.2}$$

基于 KART 理论的双层网络特性与 MLP 的网络堆叠结构的共同启发，Liu 等人[53]提出了 KAN 模型。具体而言，KART 理论可以表述为一种双层的网络结构，其内层函数将n个输入映射到(2n+1)个输出，而外层函数则将这(2n+1)个输入映射为1个输出。KAN 模型在此基础上对该双层基础结构进行了拓展，通过堆叠更多网络层来实现更深层次的特征表示，相应地，公式(2.1)可以进一步被改写为公式(2.3)：

$$f(x) = \sum_{i_{L-1}=1}^{n_{L-1}} \phi_{L-1,i_L,i_{L-1}} \left(\sum_{i_{L-2}=1}^{n_{L-2}} \cdots \sum_{i_0=1}^{n_0} \phi_{0,i_1,i_0}(x_{i_0}) \right) \tag{2.3}$$

其中n_i是 KAN 模型第i层中的节点数量。$\phi_{l,j,i}$是激活函数，它将 KAN 模型第l层的神经元i连接到 KAN 模型第$(l+1)$层的神经元j，其中$l = 0, \ldots, L-1$，$i =$

$1,\dots,n_l$以及$j=1,\dots,n_{l+1}$。

同时，具有n_l维输入和n_{l+1}维输出的 KAN 模型第l层可以用一个一维函数的矩阵来表示，其形式如公式(2.4)所示：

$$x_{l+1}=\underbrace{\begin{pmatrix}\phi_{l,1,1}(\cdot) & \phi_{l,1,2}(\cdot) & \cdots & \phi_{l,1,n_l}(\cdot)\\ \phi_{l,2,1}(\cdot) & \phi_{l,2,2}(\cdot) & \cdots & \phi_{l,2,n_l}(\cdot)\\ \vdots & \vdots & \ddots & \vdots\\ \phi_{l,n_{l+1},1}(\cdot) & \phi_{l,n_{l+1},2}(\cdot) & \cdots & \phi_{l,n_{l+1},n_l}(\cdot)\end{pmatrix}}_{\Phi_l}x_l \tag{2.4}$$

其中，Φ_l是对应于 KAN 模型第l层的函数矩阵。因此，一个典型的具有L层的 KAN 模型处理输入x以产生输出的过程如下，其形式如公式(2.5)所示：

$$KAN(x)=(\Phi_{L-1}\circ\Phi_{L-2}\circ\cdots\circ\Phi_1\circ\Phi_0)x \tag{2.5}$$

2.1.2 KAN 模型的激活函数

Liu 等人[53]将激活函数$\phi(x)$设置为基函数$b(x)$和样条函数$spline(x)$及其相应的权重矩阵w_b和w_s乘积的和，其形式如公式(2.6)所示：

$$\phi(x)=w_b b(x)+w_s spline(x) \tag{2.6}$$

基函数$b(x)$的形式如公式(2.7)所示：

$$b(x)=SiLU(x)=\frac{x}{1+e^{-x}} \tag{2.7}$$

样条函数$spline(x)$被参数化为 B 样条的线性组合，其形式如公式(2.8)所示：

$$spline(x)=\sum_i c_i B_i(x) \tag{2.8}$$

其中c_i是可训练参数。

图 2.1 展示了一个简单两层 KAN 模型的结构示意图与激活函数$\phi_{1,1,3}$的计算示例，其中网格大小 $G=3$，样条阶数 $K=3$，B 样条的数量等于 $G+K=6$。KAN 模型创新地将可学习的激活函数置于边缘，并且用样条参数化的单变量函数替换每个权重参数，从而完全消除了线性权重矩阵。这种网络架构使得 KAN 模型能够以更紧凑的模型结构实现媲美甚至超越同类模型的性能，同时显著提升模型的可解释性，为包括偏微分方程求解问题[54-56]、量子计算[57,58]、时间序列预测[59-61]

和视觉任务[62-64]在内的各种应用场景提供了一种灵活且高效的解决方案。在KAN 模型的基础上，Liu 等人[65]进一步提出了 MultKAN(Multiplicative Kolmogorov-Arnold Network)模型，该模型在加法的基础上纳入了元素乘法，以增强数据表征并捕捉更复杂的函数关系。此外，许多新型 KAN 模型的激活函数选用了经典的数学函数，特别是那些具备优异曲线拟合能力的函数。

FastKAN 模型[66]通过使用高斯径向基函数(Gaussian Radial Basis Function, GRBF)来近似三阶 B 样条，并且使用层归一化来将模型的输入保持在样条网格的范围之内，从而加快了 KAN 模型的训练速度，其激活函数的数学表达形式如公式(2.9)所示：

$$\phi(r) = e^{-\frac{r^2}{2h^2}} \tag{2.9}$$

其中$r = \|x - c\|$是输入x和中心c之间的距离，h用于控制函数的宽度。

最终，具有N个中心的 FastKAN 模型可以表示为如公式(2.10)所示：

$$FastKAN(x) = \sum_{i=1}^{N} w_i \phi(r_i) = \sum_{i=1}^{N} w_i e^{-\frac{\| x - c_i \|^2}{2h^2}} \tag{2.10}$$

其中w_i是可训练参数。

FasterKAN 模型[67]使用反射式开关激活函数(Reflectional SWitch Activation Function, RSAF)来构建激活函数，因其在具有均匀网格时易于计算的特点，在前向和后向处理速度上都要优于 FastKAN 模型。FasterKAN 模型使用的激活函数如公式(2.11)所示：

$$\phi(r) = 1 - \left(\tanh\left(\frac{r}{h} \right) \right)^2 \tag{2.11}$$

其中$r = \|x - c\|$是输入x和中心c之间的距离，h用于控制函数的宽度。

最终，具有N个中心的 FasterKAN 模型可以表示为如公式(2.12)所示：

$$FasterKAN(x) = \sum_{i=1}^{N} w_i \phi(r_i) = \sum_{i=1}^{N} w_i \left(1 - \left(\tanh\left(\frac{\| x - c_i \|}{h} \right) \right)^2 \right) \tag{2.12}$$

其中w_i是可训练参数。

Wav-KAN 模型[68]采用小波函数来构建激活函数，具体包括 Mexican hat 小波、Morlet 小波、DOG 小波以及 Shannon 小波，其数学表达形式则分别如公式

(2.13)、公式(2.14)、公式(2.15)和公式(2.16)所示：

$$\phi(x)=\frac{2}{\sqrt{3\pi}^{1/4}}(x^2-1)e^{-\frac{x^2}{2}} \tag{2.13}$$

$$\phi(x)=\cos(\omega_0 x)e^{-\frac{x^2}{2}} \tag{2.14}$$

其中 ω_0 是中心频率。

$$\phi(x)=-\frac{d}{dx}\left(e^{-\frac{x^2}{2}}\right)=x\cdot e^{-\frac{x^2}{2}} \tag{2.15}$$

$$\phi(x)=\mathrm{sinc}(x/\pi)\cdot\omega(x) \tag{2.16}$$

其中 $\omega(x)$ 是窗函数。

此外，Chebyshev KAN 模型[69]使用切比雪夫多项式来参数化激活函数，rKAN 模型[70]利用有理函数来构建激活函数，Smooth KAN 模型[71]使用结构化且平滑的嵌套函数来构建激活函数，FourierKAN-GCF 模型[72]使用傅里叶级数构建激活函数，EfficientKAN 模型[73]也利用 B 样条构建激活函数，BSRBF-KAN 模型[74]则是结合 B 样条函数与高斯径向基函数来构建激活函数。

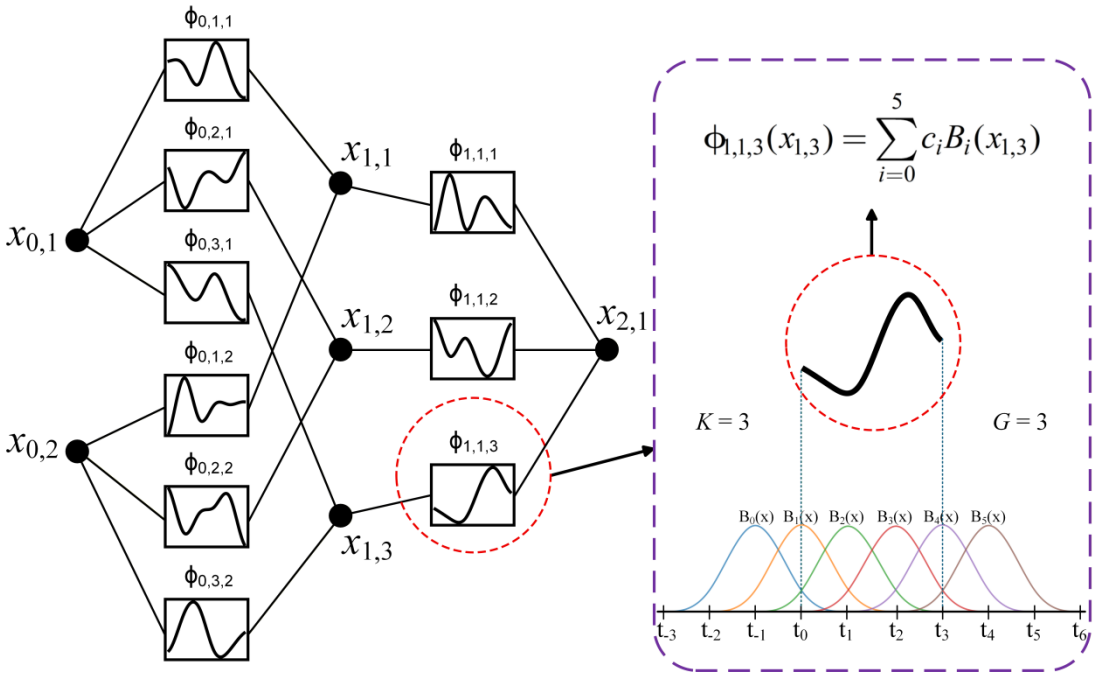


图 2.1 KAN 模型的结构示意图与激活函数$\phi_{1,1,3}$的计算示例

2.2 图像频域水印嵌入算法理论

随着模型水印技术的发展，研究者构建水印触发集的方式逐步从传统空域操作转向图像频域嵌入。空域触发集构造方法虽实现简便、计算高效，但其鲁棒性通常难以满足复杂攻击场景的需求。相较而言，频域水印依托频域变换完成信息嵌入，通过调整图像频域系数来加载水印，并利用频域特征全局关联特性，使水印均匀分布于整幅图像，从而为模型水印提供更稳定的支撑。其中，典型的频域水印实现方式主要包括离散小波变换、离散余弦变换及快速傅里叶变换等技术。

2.2.1 离散小波变换

离散小波变换(Discrete Wavelet Transform, DWT)[75]作为一种经典的时频分析方法，能够在将信号分解为不同频率成分的同时保留时间信息，从而实现时间和频率的局部化分析。将其应用于图像频域水印嵌入时，首先将原始图像转换为灰度图像，再利用选定的小波基函数对灰度图像进行分解，就能得到四个子带分量：低频近似分量（LL）以及水平（HL）、垂直（LH）、对角（HH）三个方向的高频细节分量。对于尺寸为$M \times N$的图像$I(x,y)$，其近似系数$W_{\varphi}(j_0,m,n)$和细节系数$W_{\psi}^{i}(j,m,n)$可分别通过图像与尺度函数$\varphi_{j_0,m,n}(x,y)$和小波函数$\psi_{j,m,n}^{i}(x,y)$进行内积运算得到，表达式如公式(2.17)和(2.18)所示：

$$W_{\varphi}(j_0,m,n)=\frac{1}{\sqrt{MN}}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}I(x,y)\cdot\varphi_{j_0,m,n}(x,y) \tag{2.17}$$

$$W_{\psi}^{i}(j,m,n)=\frac{1}{\sqrt{MN}}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}I(x,y)\cdot\psi_{j,m,n}^{i}(x,y) \tag{2.18}$$

其中$i\in\{H,V,D\}$分别对应水平、垂直及对角方向，j是分解尺度，j_0是初始分解尺度。最后将水印信息嵌入选定的系数中，并通过逆离散小波变换即可得到水印图像，其逆变换公式如(2.19)所示：

$$\begin{aligned}I(x,y)=&\frac{1}{\sqrt{MN}}\sum_{m}\sum_{n}W_{\varphi}(j_0,m,n)\varphi_{j_0,m,n}(x,y)\\&+\frac{1}{\sqrt{MN}}\sum_{j=j_0}^{J}\sum_{m}\sum_{n}\sum_{i\in\{H,V,D\}}W_{\psi}^{i}(j,m,n)\psi_{j,m,n}^{i}(x,y)\end{aligned} \tag{2.19}$$

2.2.2 离散余弦变换

离散余弦变换(Discrete Cosine Transform, DCT)[76]作为一种经典的频域分析方法，能够将信号从时域转换至频域，实现信号能量的重新分布与压缩表示。将其应用于图像频域水印嵌入时，首先将原始图像转换为灰度图像，再对灰度图像进行 DCT，得到对应的频域系数矩阵。对于尺寸为$M \times N$的图像$I(x,y)$，其 DCT 变换系数$F(u,v)$可由公式(2.20)计算得到：

$$F(u,v)=\frac{2}{\sqrt{MN}}C(u)C(v)\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}I(x,y)\cos\frac{(2x+1)u\pi}{2M}\cos\frac{(2y+1)v\pi}{2N} \quad (2.20)$$

其中$C(\xi)=\begin{cases}\frac{1}{\sqrt{2}}, & \xi=0\\ 1, & \xi\neq 0\end{cases}$。最后将水印信息嵌入选定的系数中，并通过逆离散余弦变换即可得到水印图像，其逆变换公式如(2.21)所示：

$$I(x,y)=\frac{2}{\sqrt{MN}}\sum_{u=0}^{M-1}\sum_{v=0}^{N-1}C(u)C(v)F(u,v)\cos\frac{(2x+1)u\pi}{2M}\cos\frac{(2y+1)v\pi}{2N} \quad (2.21)$$

2.2.3 快速傅里叶变换

快速傅里叶变换(Fast Fourier Transform, FFT)[77]作为一种高效的频域分析方法，能够将信号从时域快速转换至频域，从而揭示其频率构成与相位信息。将其应用于图像频域水印嵌入时，首先将原始图像转换为灰度图像，然后再对灰度图像进行 FFT，得到由幅度谱和相位谱组成的频域表示。对于尺寸为$M \times N$的图像$I(x,y)$，其 FFT 变换系数$F(u,v)$可由公式(2.22)计算得到：

$$F(u,v)=\frac{1}{\sqrt{MN}}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}I(x,y)\cdot e^{-j2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)} \quad (2.22)$$

其中j为虚数单位。最后将水印信息嵌入选定的系数中，并通过逆快速傅里叶变换即可得到水印图像，其逆变换公式如(2.23)所示：

$$I(x,y)=\frac{1}{\sqrt{MN}}\sum_{u=0}^{M-1}\sum_{v=0}^{N-1}F(u,v)\cdot e^{j2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)} \quad (2.23)$$

2.3 模型水印技术基础

2.3.1 模型水印分类

针对图像分类任务，KAN 模型的版权保护可借助模型水印技术实现，其流程主要划分为水印嵌入与水印验证两大关键环节。在嵌入环节，核心在于将具备可验证性的版权信息隐秘植入模型中，同时维持模型在分类任务上的精度与性能的稳定，避免对原有功能造成显著干扰。在验证环节，模型所有者通过预设的验证策略与交互协议，从可疑模型中还原出水印信息，进而完成模型所有权的核验与确认。依据验证阶段对模型内部细节的访问权限，相关水印方案可分为白盒水印与黑盒水印，前者需要获取模型内部细节，而后者则仅凭借模型的输入输出响应即可完成验证。

白盒水印方法通常采用参数空间映射的方式实现版权水印信息的嵌入[17]。具体而言，首先从模型参数全集θ中选定一个特定子集$\theta_B \subset \theta$来作为水印的承载区域，然后利用密钥K对原始二进制水印序列B进行加密变换，以提升水印信息的安全性与抗攻击能力，最后通过联合优化目标模型参数，在保持模型原有任务性能的前提下完成水印信息的嵌入，该过程可形式化表示为公式(2.24)：

$$\tilde{\theta} = \arg\min_{\theta}\left[\mathcal{L}_{\text{task}}(f(X;\theta),Y) + \lambda\mathcal{L}_{\text{wm}}(\sigma(K\cdot\theta_B),B)\right] \tag{2.24}$$

其中X是输入数据，Y是目标标签集，$\sigma(\cdot)$是符号函数，$\mathcal{L}_{task}$是原始任务损失函数，$\mathcal{L}_{wm}$是水印任务损失函数，λ是平衡两项损失的超参数。

在版权验证阶段，模型所有者需获取可疑模型的内部信息，明确水印嵌入的具体参数θ_B，并持有正确的加密密钥K，才能完成水印的提取。通过计算密钥K与参数θ_B的点积并经符号函数$\sigma(\cdot)$量化即可提取出相应的二进制序列，最后将其与原始水印信息B进行匹配度比对便可确认模型的版权归属。

相比于白盒水印方法，黑盒水印的基本原理在于构建具备特定语义特征的触发样本集[27]，并利用协同训练策略来将水印信息嵌入到模型之中。对于标准图像分类任务，目标是学习映射函数 $f_\theta : \mathcal{X} \to \mathcal{Y}$，其中 $\mathcal{X}$ 是输入域，$\mathcal{Y}$ 是目标类的集合。学习过程涉及从训练数据集中获得最佳参数 θ。在黑盒水印的背景下，模

型所有者的目标是将水印嵌入 KAN 模型中，以便通过与模型的交互来验证水印模型的所有权。水印嵌入过程涉及到使用原始训练集和触发集来训练模型。一般来说，将干净的训练样本 $(\mathcal{X}_i,\mathcal{Y}_i)$ 转换为由触发集图像构造器 $\mathcal{G}$ 生成的触发集样本 $(\mathcal{G}(\mathcal{X}_i),\mathcal{Y}_t)$，其中 $\mathcal{Y}_t$ 表示用于触发集图像分类的预定义标签。然后，函数 f_θ 的训练基于干净样本和触发集样本，可以定义为公式(2.25)：

$$\begin{cases} f_\theta(\mathcal{X}_i)=\mathcal{Y}_i \\ f_\theta(\mathcal{G}(\mathcal{X}_i))=\mathcal{Y}_t \end{cases} \tag{2.25}$$

利用此过程，触发集图像构造器 $\mathcal{G}$ 可以将干净图像 $\mathcal{X}$ 转换为触发集图像 $\mathcal{G}(\mathcal{X})$。随后，带水印的 KAN 模型将 $\mathcal{G}(\mathcal{X})$ 分类为特定的标签 $\mathcal{Y}_t$，同时将干净图像 $\mathcal{X}$ 分类为正常类别。

2.3.2 常见的模型水印攻击方法

在模型版权保护场景中，攻击者为非法侵占模型的所有权或破坏水印的保护效果，提出了多种模型水印攻击方法。这些方法基于不同的技术原理，针对水印嵌入特性实施攻击，旨在擦除水印信息或混淆版权归属。下面将对几类常见的模型水印攻击方法进行详细说明。

剪枝攻击是一种利用模型剪枝技术[78]实施的模型水印攻击方式，其攻击逻辑源于模型剪枝的核心特性。模型剪枝是模型压缩领域的关键技术之一，其核心目标是在尽可能维持原始模型性能的前提下，通过去除网络中的冗余参数及对最终输出贡献较小的连接，降低模型的计算复杂度与存储开销，从而提升模型的推理效率。主流的剪枝方式可归纳为以下几类：通道剪枝通过评估各卷积通道对模型输出的贡献程度，筛选并移除作用微弱的通道，通常基于参数重要性或激活值的统计特性完成判别，在有效削减冗余参数的同时，最大程度保留模型精度并减轻计算负担；权重剪枝采用基本的参数级剪枝策略，根据权重绝对值的大小或梯度信息去除网络中权重较低或贡献较低的连接，可通过权重置零或直接删除的方式实现模型轻量化，进而降低计算资源消耗；结构化剪枝则面向更高层级的网络单元，以完整网络层、功能模块或神经元为单位进行裁剪，依据不同结构对模型性能的影响程度进行筛选，能够显著降低模型参数量与计算复杂度。剪枝攻击基于

上述模型剪枝方式，通过删除模型中对于原始任务冗余的水印任务相关参数来擦除水印信息。然而，常规剪枝方法通常难以精确区分原始任务参数与水印任务参数，随着剪枝强度的提升，两类任务性能往往同步下降。因此，攻击者需在维持原任务性能与削弱水印效果之间寻求平衡。

微调攻击是一种针对模型水印的常见攻击方式，其技术核心源自于模型微调技术[17]。模型微调技术作为基于预训练模型进行任务适配的常用优化手段，其核心是在不改变模型原有结构的基础上，利用新的数据集对模型的参数进行进一步训练，使其适配目标任务的分布特征，并且通过梯度更新对模型的权重进行迭代调整，可以充分继承预训练模型的通用特征与强泛化能力，让模型能够快速收敛至新场景最优状态，微调过程中可以采用全量微调以及参数高效微调（如逐层解冻）等不同策略，这也为微调攻击的实施提供了技术路径。微调攻击正是利用了这些特性，其核心实施逻辑是不改变水印模型的原有结构，采用干净数据集对水印模型进行再训练，由于水印模型的参数空间中通常包含有原始任务与水印任务这两个核心任务，而微调攻击的关键在于再训练过程仅使用与原始任务相关的干净数据、不引入任何与水印任务相关的信息，因此经过若干轮迭代训练后，模型的参数空间会在微调机制的作用下发生适应性调整，模型会逐步强化对原始任务的学习与适配，原本嵌入其中的水印任务则是会被逐渐遗忘，最终在保持原始任务性能不变甚至有所增强的同时擦除水印信息，导致版权所有者无法通过水印验证确认模型归属。

覆盖攻击作为一种针对模型水印的典型攻击手段，攻击者会在已嵌入原始水印的模型中重新植入新的水印信息[79]，通过新水印的干扰作用破坏原有水印的有效性，进而降低旧水印的提取率。与此同时，攻击者可通过提取自身植入的新水印，虚假宣称对该模型拥有所有权，最终混淆模型的真实版权归属，导致原版权所有者无法通过水印验证确认其合法权益。

此外，除了单一攻击方式外，攻击者还可以采用组合攻击的手段来去除模型水印，其中剪枝-微调攻击[80]是较为典型的一种攻击方式，即先对水印模型实施剪枝攻击，再进一步开展微调攻击，通过组合多种水印攻击方式，实现模型水印的高效去除。

2.3.3 模型水印的评价指标

基于图像分类任务的模型水印设计需要兼顾水印嵌入前后模型原始任务性能的稳定性、水印自身嵌入的稳定性以及后续检测的有效性，其设计合理性与保护有效性通常从保真度、鲁棒性和隐蔽性三个维度进行衡量，具体定义如下：

（1）保真度：核心包含两方面要求，一是任务保真度，即水印嵌入过程中需尽可能不影响目标模型的原始任务性能。由于水印信息通过模型的权重进行存储，需确保嵌入操作完成后，模型仍能稳定执行预定的图像分类任务，将水印对模型性能的负面影响降至最低；二是水印保真度，即要求水印成功嵌入后，模型所有者能够准确提取嵌入的水印信息。尤其在模型遭遇非法传播时，模型所有者能够以较高概率从水印模型中恢复水印，从而有效确认模型版权归属。

（2）鲁棒性：要求水印模型在遭遇水印去除攻击后，依然能够有效地进行模型版权验证。模型微调、剪枝等技术本用于优化神经网络的学习效率与推理速度，但也可能被攻击者恶意利用来篡改模型参数或破坏水印完整性。因此，水印模型必须具备较高的鲁棒性，确保水印信息在遭遇攻击后不易被删除或篡改，从而维持版权验证的有效性。

（3）隐蔽性：要求模型水印需具备良好的隐蔽性以尽可能降低对模型自身的影响。若隐蔽性不足，攻击者可能通过尝试检测水印判断出模型水印方案及其嵌入位置，进而可能通过伪造水印宣称模型所有权或通过篡改模型参数干扰水印验证过程。

2.4 本章小结

本章系统阐述了面向图像分类任务的 KAN 模型水印的相关技术基础，主要涵盖 KAN 模型基础概念、图像频域水印嵌入算法理论以及模型水印技术基础三个核心部分。首先介绍了 KAN 模型的理论根基，同时分析其网络结构特性，包括置于“边”上的可学习激活函数、样条参数化设计以及多层复合结构，并且梳理了 FastKAN、FasterKAN 等变体 KAN 模型在激活函数上的创新，为后续设计适用于 KAN 模型的水印嵌入方案提供基础。随后详细阐述了离散小波变换、离

散余弦变换及快速傅里叶变换三种频域嵌入方法的数学原理，并且分析了各类变换嵌入水印信息的实现机制，为构造面向 KAN 模型的水印触发集提供了技术支撑。最后系统梳理了白盒水印与黑盒水印的分类体系以及形式化定义，分析了剪枝攻击、微调攻击、覆盖攻击及组合攻击等常见水印攻击方法的技术机理，并且明确了保真度、鲁棒性和隐蔽性三大评价指标的具体内涵，为后续构建面向 KAN 模型的水印方案提供理论依据。

第三章　基于频域扰动和交替训练的 KAN 模型水印

3.1 引言

随着人工智能技术的快速发展，各类新型神经网络架构不断涌现，并在计算机视觉[2]和自然语言处理[3]等众多领域取得了显著成效。作为一种新兴的神经网络架构，KAN 模型[53]创新性地将可学习的激活函数置于网络的“边”上，而非像多层感知机那样在“节点”上使用固定激活函数，显著增强了模型拟合复杂函数的能力。这一设计不仅为求解偏微分方程、量子计算、时间序列预测等任务提供了更有效的工具，也为神经网络架构的发展开辟出了新的方向。然而，随着 KAN 模型在实际应用中的广泛部署，其面临的未经授权访问与滥用风险日益突出。未经授权的第三方可以轻易从盗版模型中窃取有价值信息，进而引发隐私泄露、版权侵权、恶意篡改等一系列安全问题，使得 KAN 模型的知识产权保护面临严峻挑战。在此背景下，在不影响 KAN 模型原有功能的前提下有效嵌入水印信息以验证 KAN 模型所有权，已成为一个迫切需要解决的关键问题。

目前，主流的黑盒水印方法虽然在一定程度上能够实现模型所有权验证，但应用于 KAN 模型时仍存在局限。由于 KAN 模型的激活函数善于拟合连续平滑的函数，其在训练过程中会优先拟合样本量大、损失下降明显的原始任务，导致水印任务的学习能力存在不足。黑盒水印的嵌入方法主要分为训练时嵌入(train-to-embed)和微调时嵌入(fine-tune-to-embed)两种。训练时嵌入是将原始训练集与触发集合并为新的训练集来训练目标模型以实现水印嵌入；微调时嵌入则是在已经使用原始训练集进行预训练的目标模型的基础上，使用触发集进行微调来完成水印嵌入。遗憾的是，这两种嵌入方法在应用于 KAN 模型时均未能取得满意效果。具体而言，训练时嵌入策略难以使 KAN 模型充分学习水印任务；微调时嵌入策略则会导致 KAN 模型在原始任务上的性能显著下降。

针对上述挑战，本章提出了一种基于频域扰动和交替训练的鲁棒黑盒 KAN 模型水印方法。该方法通过在频域添加扰动，使触发信号分布并且影响样本的整个空域，从而让模型更好地学习到触发模式。同时，为了增强模型在训练过程中

学习水印任务的能力，该方案采用交替训练的方式优化模型参数，并引入模型参数扰动损失以模拟攻击所引发的参数变化。实验结果表明，本章方法能够在不影响 KAN 模型原始任务性能的前提下，实现 100%的水印任务分类准确率（即使在 50%剪枝率下），并在微调攻击后保持 98.67%以上的平均性能。

3.2 基于频域扰动和交替训练的 KAN 模型水印方案

3.2.1 总体框架

图 3.1 展示了基于频域扰动和交替训练的 KAN 模型水印方案总体框架，主要由触发集构建、水印嵌入和所有权验证三个阶段构成，整个方案基于交替训练策略和参数扰动机制实现 KAN 模型的黑盒水印保护。

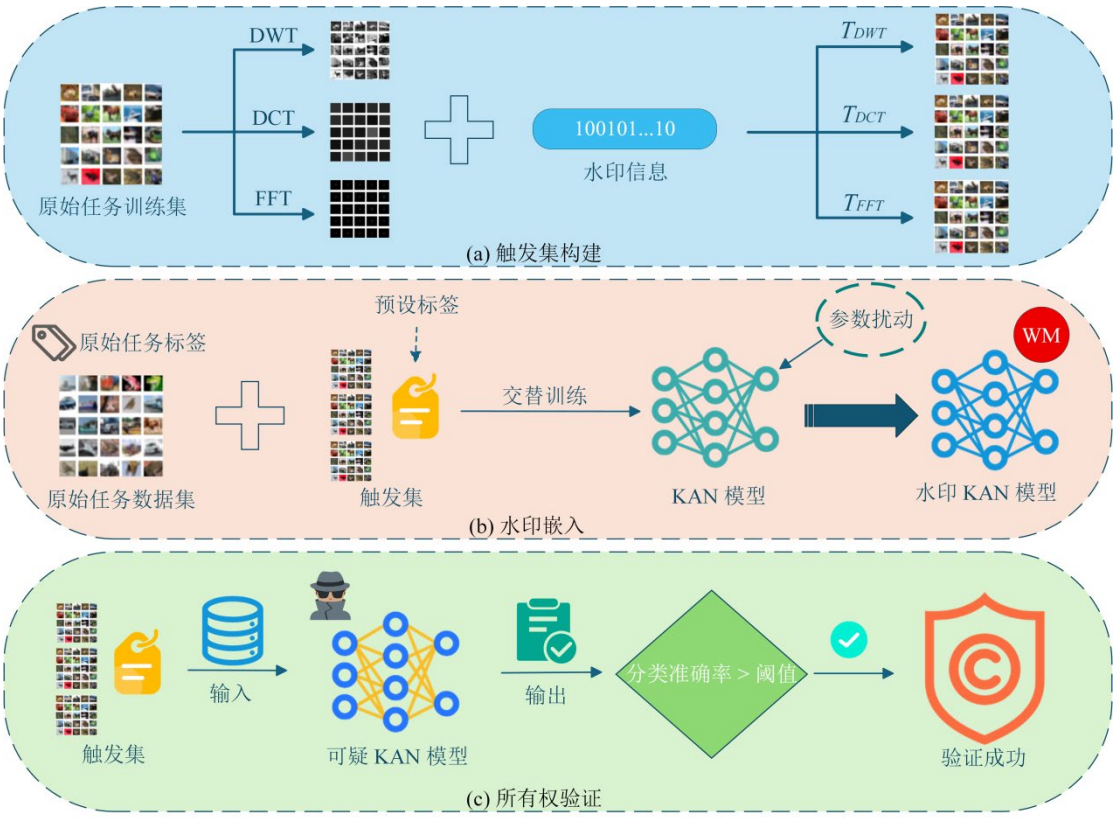


图 3.1 基于频域扰动和交替训练的 KAN 模型水印框架

在触发集构建阶段，使用 DWT、DCT 和 FFT 三种变换域技术，将水印信息嵌入到干净样本中，分别生成触发集 T_{DWT}、T_{DCT} 和 T_{FFT}。在水印嵌入阶段，将原始训练集与触发集交替输入 KAN 模型进行训练，原始训练集样本保留其原始标

签，触发集样本被赋予特定的预定义标签；同时，在训练过程中引入模型参数扰动损失，通过对模型参数施加高斯噪声来模拟水印去除攻击从而增强水印的鲁棒性，最终获得含水印的 KAN 模型。在所有权验证阶段，将可疑 KAN 模型通过黑盒方式进行查询，输入构建的触发集并获取其分类输出，若触发集分类准确率超过预定阈值，则表明水印验证成功，确认该可疑模型为盗版模型从而完成模型的所有权认证。

3.2.2 触发集构建

在频域中嵌入水印信息可以显著增强水印的鲁棒性，因为在频域中对水印的微小扰动会在空间域中分散到更广泛的区域，从而使模型能够更有效地学习水印特征，从而更好地满足现实场景中对水印技术的严格要求。基于不同的频域变换方式，本章提出了一种用于 KAN 模型所有权验证的黑盒水印方法，该方法通过将水印信息嵌入到干净样本的不同变换域中来构建触发集。

DWT 域嵌入：作为一项经典且高效的信号处理技术，DWT 在数字图像水印领域得到了广泛应用。在本章中，我们选择 Haar 小波基对无水印的原始图像进行 DWT，从而得到四个不同的子带：LL、HL、LH 和 HH。我们选择 LL 子带作为水印嵌入域，并以 0.5 的嵌入强度嵌入水印信息，从而生成带水印的 LL'。最后，我们对 LL'、HL、LH 和 HH 进行逆 DWT，以构建触发集 T_{DWT}。

DCT 域嵌入：由于其出色的正交变换特性，DCT 在图像水印的嵌入和提取中发挥着重要作用。在本章中，我们对无水印的原始图像进行 DCT，以嵌入强度为 0.5 的水印信息，然后执行逆 DCT 以构建触发集 T_{DCT}。

FFT 域嵌入：作为计算离散傅里叶变换(Discrete Fourier Transform, DFT）及其逆变换的高效算法，FFT 在数字图像水印领域发挥着至关重要的作用。在本章中，我们对未嵌入水印的原始图像执行 FFT，随后以 0.5 的嵌入强度嵌入水印信息。最后，我们执行逆 FFT 以构建触发集 T_{FFT}。

3.2.3 水印嵌入

交替训练策略的动机在于，需要将水印任务与 KAN 模型的原始训练任务有效融合。传统的训练方法，如训练时嵌入和微调时嵌入，在鲁棒性和性能方面存在局限性。交替训练策略通过平衡特征学习来克服这些局限性。在模型训练的过

程中，交替使用原始训练集和触发集，KAN 模型能够更好地将触发集的特征与原始训练集的特征相结合，从而确保模型在有效学习水印任务的同时，不会显著降低其在原始任务上的性能。此外，我们在训练过程中引入了一种模型参数扰动损失，通过向模型的参数施加扰动来模拟水印去除攻击，从而增强了水印的鲁棒性，使模型更能抵御常见的水印去除攻击。

具体而言，在水印嵌入阶段，原始训练集和触发集在 KAN 模型的训练过程中交替使用，以确保 KAN 模型能够有效学习触发集的特征，同时保持其在原始任务上的分类性能。原始训练集中的样本保留其原始标签，而触发集中的样本则被分配一个特定的标签。鉴于记忆水印信息的权重可能会在水印去除攻击的影响下发生改变，从而降低 KAN 模型在水印任务上的分类准确性，所提出的方法在水印嵌入阶段引入了模型参数扰动损失，通过扰动模型的参数来模拟水印去除攻击，从而增强水印的鲁棒性。为此，我们需要确定模型参数扰动损失中扰动的大小。扰动后的模型参数$w_{l,i}^{p}$定义为公式(3.1)：

$$w_{l,i}^{p} = w_{l,i} + \alpha_l X_{l,i}; \quad X_{l,i} \sim N(0,\sigma_l^2) \tag{3.1}$$

其中，$w_{l,i}$表示扰动前 KAN 模型第 l 层的第 i 个模型参数。α_l是表示注入噪声$X_{l,i}$大小的系数。$X_{l,i}$是从均值为零、方差为σ_l^2的高斯分布中采样的噪声项，σ_l^2是扰动前 KAN 模型第 l 层参数的标准差。

随后，模型训练的总损失 L 定义为公式(3.2)：

$$L = L_{CE} + \kappa L_{CE}^{p} \tag{3.2}$$

其中 L_{CE} 表示交叉熵损失，κ 是一个可调参数，L_{CE}^{p} 表示模型参数扰动后水印任务的交叉熵损失。

3.2.4 所有权验证

本章提出的方法采用黑盒方式进行验证。在黑盒场景中，模型所有者A拥有一个为特定任务T设计的深度模型M，并利用该模型提供远程服务S。可能存在模型攻击者B，其目标是非法窃取A的合法模型 M，以构建一个用于与原始任务T高度相似的任务 T'的盗版模型 M'，并且建立新的远程服务 S'，以获取非法利润。在此情况下，模型的所有权验证要求模型所有者 A 查询可疑 KAN 模型服务 S'在

触发集上的分类结果，以验证泄露版本 *M'*的所有权。在实际应用中，如果可疑 KAN 模型服务 *S'*在触发集上的分类准确率超过某个阈值，则可以证明模型 *M'*是盗版模型。

3.3 实验结果与分析

3.3.1 实验设置

实验选用的图像数据集包括 MNIST[81]、CIFAR-10[82]以及 CIFAR-100[82]。其中，MNIST 数据集包含 70000 张灰度图像，每张图像的大小为 28×28，分为 10 个类别；CIFAR-10 数据集由 60000 张彩色图像组成，每张图像的大小为 32×32×3，分为 10 个类别；CIFAR-100 数据集包含 60000 张彩色图像，每张图像的大小为 32×32×3，分为 100 个类别。在本章中，使用 EfficientKAN 模型、FastKAN 模型和 FasterKAN 模型作为目标 KAN 模型。表 3.1 展示了在不同数据集上训练的 KAN 模型的配置。

表 3.1 不同数据集上训练的 KAN 模型的配置

数据集	模型	网络结构	可学习的激活函数
MNIST	EfficientKAN	[784, 64, 10]	B 样条函数
	FastKAN	[784, 64, 10]	高斯径向基函数
	FasterKAN	[784, 64, 10]	反射式开关激活函数
CIFAR-10	EfficientKAN	[3072, 256, 10]	B 样条函数
	FastKAN	[3072, 256, 10]	高斯径向基函数
	FasterKAN	[3072, 256, 10]	反射式开关激活函数
CIFAR-100	EfficientKAN	[3072, 256, 100]	B 样条函数
	FastKAN	[3072, 256, 100]	高斯径向基函数
	FasterKAN	[3072, 256, 100]	反射式开关激活函数

在不失一般性的前提下，本章提出的方法与训练时嵌入和微调时嵌入这两种模型水印嵌入的训练策略进行了比较。训练时嵌入是将原始训练集和触发集合并成一个新的训练集来训练目标 KAN 模型，从而实现水印的嵌入。微调时嵌入则

是在已经在原始训练集上预训练的 KAN 模型上使用触发集来进行微调，从而实现水印的嵌入。

在使用本章提出的方法训练 KAN 模型时，我们使用 AdamW 优化器来进行模型参数的优化，其中初始学习率设置为10^{-3}，权重衰减设置为10^{-4}。学习率通过指数调度器进行调整，其中 gamma 设置为 0.8，调度器从模型训练第一个周期开始就生效。我们将批量大小设置为 100，并对每个 KAN 模型进行 45 个周期的训练。当应用微调嵌入时，会进行一个额外的五个周期的训练阶段，初始学习率设置为10^{-4}。此外，可调参数的设置分别为α_l=0.01和κ=0.3。所有实验均在配备 NVIDIA GeForce RTX 4060 GPU 和 Intel(R) Core(TM) i7-14650HX CPU 以及 16.0 GB 内存的计算机环境中进行。

3.3.2 保真度分析

任务保真度要求 KAN 模型在嵌入水印后，其原始任务的分类性能不能出现明显的下降。表 3.2 展示了不同 KAN 模型在嵌入水印前在原始任务上的分类准确率。表 3.3 展示了使用不同触发集和训练策略嵌入水印后，不同 KAN 模型在原始任务上的分类准确率。从表 3.2 可以看出，由于学习能力不同，不同 KAN 模型在不同的图像分类任务中的表现各异。从表 3.3 可以看出，与训练时嵌入和本章提出的方法相比，微调时嵌入会导致 KAN 模型在原始任务上的分类准确率出现明显下降。通过对比表 3.2 和表 3.3，我们可以观察到，在使用训练时嵌入和本章提出的方法嵌入水印后，KAN 模型在原始任务上的分类性能并没有出现显著的减弱，甚至在某些情况下还有所提升，这表明本章提出的方法在实际应用中具有巨大的潜力。

表 3.2 水印嵌入前 KAN 模型在原始任务上的分类准确率

模型	数据集		
	MNIST (%)	CIFAR-10 (%)	CIFAR-100 (%)
EfficientKAN	96.97	56.85	29.40
FastKAN	97.53	53.39	22.87
FasterKAN	97.60	54.43	26.63

表 3.3 不同触发集和训练策略下 KAN 模型在原始任务上的分类准确率

数据集	模型	训练策略	触发集		
			T_{DWT}(%)	T_{DCT}(%)	T_{FFT}(%)
MNIST	EfficientKAN	训练时嵌入	97.05	96.94	97.01
		微调时嵌入	90.78	87.47	96.72
		本章方法	97.41	97.68	97.49
	FastKAN	训练时嵌入	97.46	97.45	97.34
		微调时嵌入	85.31	86.98	87.11
		本章方法	97.69	97.69	97.75
	FasterKAN	训练时嵌入	97.61	97.68	97.62
		微调时嵌入	89.07	89.62	91.60
		本章方法	97.67	97.72	97.67
CIFAR-10	EfficientKAN	训练时嵌入	56.81	56.80	56.89
		微调时嵌入	46.51	47.12	50.18
		本章方法	56.39	56.76	57.25
	FastKAN	训练时嵌入	53.22	53.02	52.64
		微调时嵌入	43.04	44.05	45.54
		本章方法	54.70	54.09	55.28
	FasterKAN	训练时嵌入	54.13	54.40	53.92
		微调时嵌入	45.93	46.26	48.91
		本章方法	53.55	53.15	53.19
CIFAR-100	EfficientKAN	训练时嵌入	29.34	29.84	29.77
		微调时嵌入	20.97	21.24	23.78
		本章方法	27.69	28.03	27.51
	FastKAN	训练时嵌入	23.40	23.05	23.17
		微调时嵌入	18.41	19.58	20.60
		本章方法	24.81	24.32	24.75
	FasterKAN	训练时嵌入	26.83	26.65	26.36
		微调时嵌入	23.43	23.59	23.64
		本章方法	26.70	25.60	26.37

此外，水印保真度要求嵌入的水印信息能够被准确重构。与零比特水印策略一致，我们的目标是在 KAN 模型中检测水印的存在。我们比较了不同 KAN 模型在基于不同变换域构建的触发集上，采用不同水印嵌入训练策略时水印任务的分类准确性。如表 3.4 所示，可以看出，即使在不同的数据集和基于不同变换域构建的触发集上，与训练时嵌入和微调时嵌入相比，本章提出的方法在将水印嵌入 KAN 模型后，水印任务的分类准确性仍能达到 100%，这证明了所提出的方法的通用性和优越性。

表 3.4 不同触发集和训练策略下 KAN 模型在水印任务上的分类准确率

数据集	模型	训练策略	触发集		
			T_{DWT}(%)	T_{DCT}(%)	T_{FFT}(%)
MNIST	EfficientKAN	训练时嵌入	100	100	100
		微调时嵌入	43.00	14.00	100
		本章方法	100	100	100
	FastKAN	训练时嵌入	100	100	100
		微调时嵌入	12.00	19.00	0
		本章方法	100	100	100
	FasterKAN	训练时嵌入	80.00	96.00	100
		微调时嵌入	25.00	13.00	100
		本章方法	100	100	100
CIFAR-10	EfficientKAN	训练时嵌入	82.00	100	93.00
		微调时嵌入	62.00	52.00	74.00
		本章方法	100	100	100
	FastKAN	训练时嵌入	100	100	100
		微调时嵌入	76.00	67.00	81.00
		本章方法	100	100	100
	FasterKAN	训练时嵌入	58.00	78.00	75.00
		微调时嵌入	54.00	51.00	48.00
		本章方法	100	100	100
CIFAR-100	EfficientKAN	训练时嵌入	82.00	100	93.00
		微调时嵌入	54.00	47.00	67.00
		本章方法	100	100	100
	FastKAN	训练时嵌入	100	100	100
		微调时嵌入	53.00	65.00	57.00
		本章方法	100	100	100
	FasterKAN	训练时嵌入	35.00	77.00	77.00
		微调时嵌入	26.00	43.00	31.00
		本章方法	100	100	100

最后，考虑到黑盒水印是利用模型的冗余性来嵌入后门水印信息，为了评估本章提出的方法对 KAN 模型在原始任务上分类性能的影响，我们在 MNIST 数据集上进行了实验，并且绘制了水印嵌入过程中 KAN 模型在原始任务训练集和验证集上的分类准确率曲线和损失曲线，如图 3.2 所示。实验结果表明，训练集和验证集的准确率随着训练轮次稳步提升，而相应的损失则是不断下降，且验证损失没有出现任何反弹。这表明水印嵌入过程不会损害 KAN 模型在原始任务上的泛化能力。

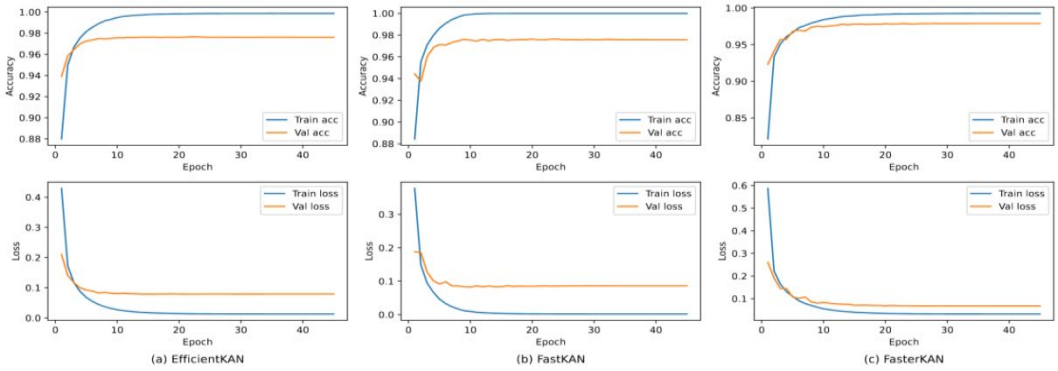
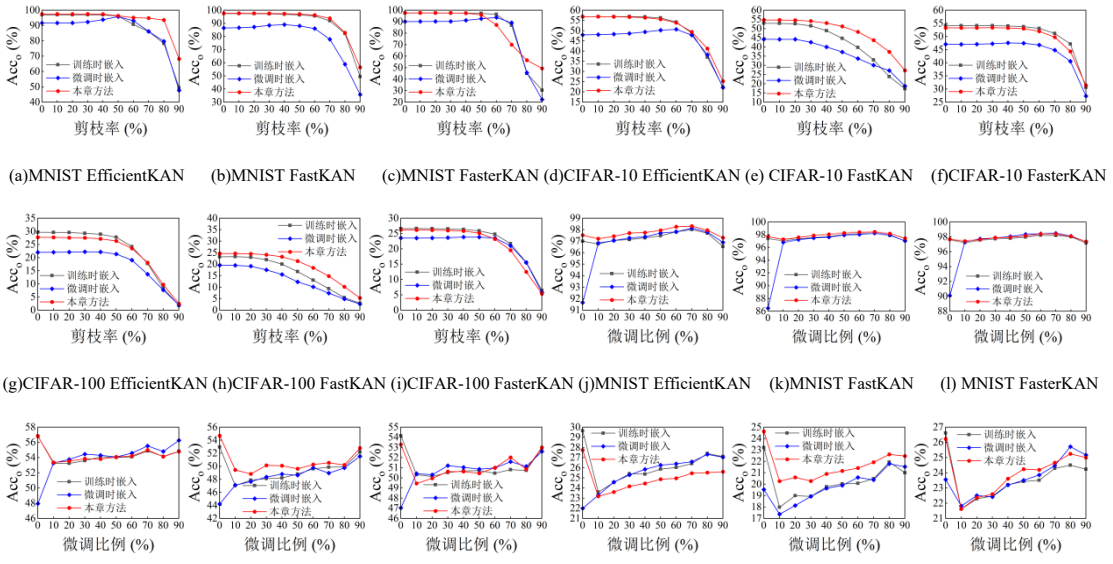


图 3.2 在 MNIST 数据集上，对(a)EfficientKAN、(b)FastKAN 和(c)FasterKAN 进行水印嵌入时的训练和验证曲线

3.3.3 鲁棒性分析

为了评估模型水印的鲁棒性，我们比较了两种最常用的应用于模型的水印去除攻击：剪枝攻击和微调攻击。对于剪枝攻击，我们对水印 KAN 模型应用了L_1范数剪枝策略，其中剪枝率表示被剪枝参数的百分比。对于微调攻击，我们使用不同比例的验证集对水印 KAN 模型进行微调，其中微调比例表示用于微调的验证集的比例。我们评估了上述攻击下 KAN 模型在三个不同的触发集T_{DWT}、T_{DCT}和T_{FFT}上，对原始任务和水印任务的平均性能，结果如图 3.3 和图 3.4 所示。我们定义了Acc_o和Acc_w，其中Acc_o表示 KAN 模型在原始任务上的分类准确率，而Acc_w则表示 KAN 模型在水印任务上的分类准确率。

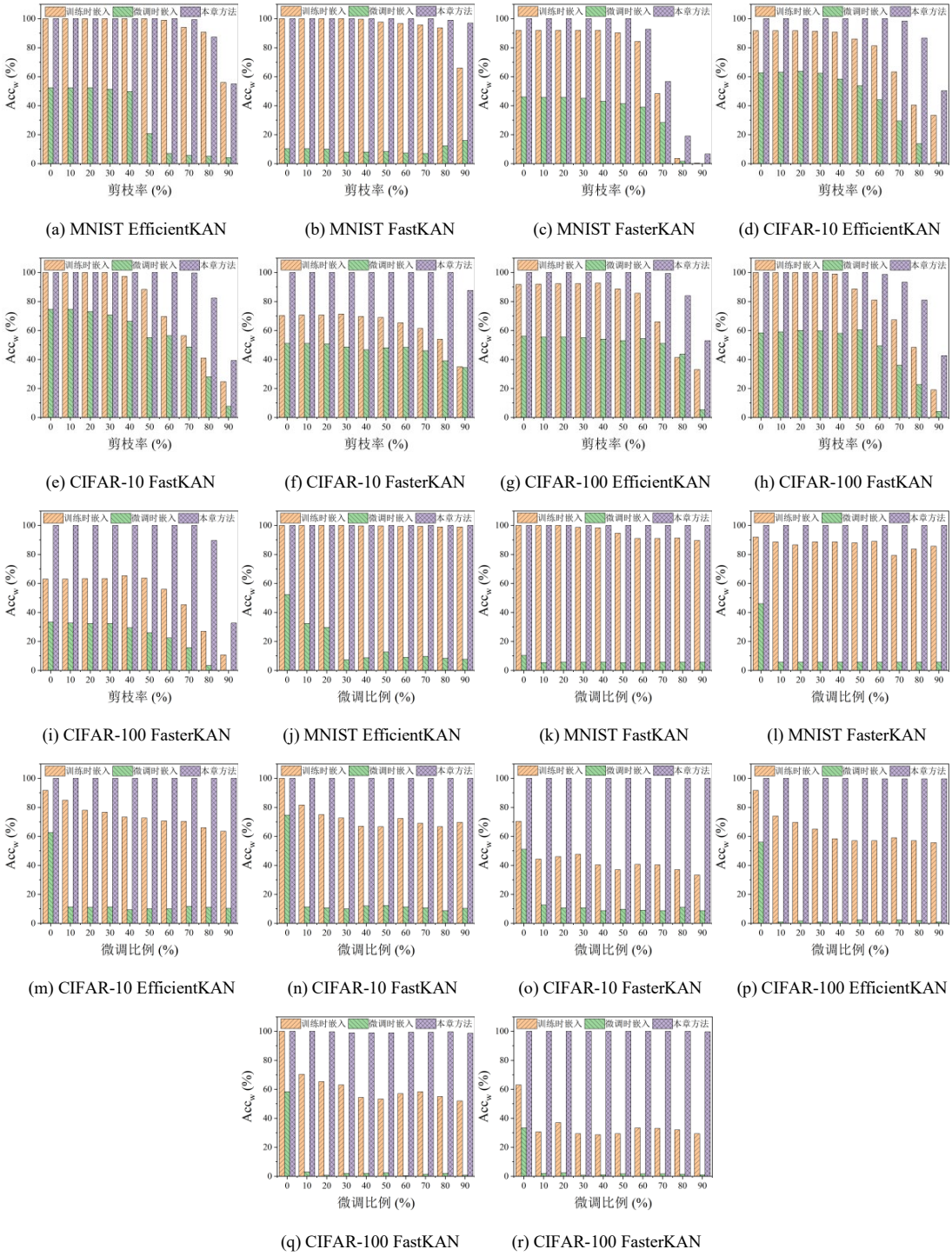


(a)MNIST EfficientKAN (b)MNIST FastKAN (c)MNIST FasterKAN (d)CIFAR-10 EfficientKAN (e) CIFAR-10 FastKAN (f)CIFAR-10 FasterKAN

(g)CIFAR-100 EfficientKAN (h)CIFAR-100 FastKAN (i)CIFAR-100 FasterKAN (j)MNIST EfficientKAN (k)MNIST FastKAN (l) MNIST FasterKAN

(m)CIFAR-10 EfficientKAN(n)CIFAR-10 FastKAN(o)CIFAR-10 FasterKAN (p)CIFAR-100 EfficientKAN(q)CIFAR-100 FastKAN(r)CIFAR-100 FasterKAN

图 3.3 原始任务上的分类准确率：(a)-(i)剪枝攻击后性能评估；(j)-(r)微调攻击后性能评估



(a) MNIST EfficientKAN (b) MNIST FastKAN (c) MNIST FasterKAN (d) CIFAR-10 EfficientKAN

(e) CIFAR-10 FastKAN (f) CIFAR-10 FasterKAN (g) CIFAR-100 EfficientKAN (h) CIFAR-100 FastKAN

(i) CIFAR-100 FasterKAN (j) MNIST EfficientKAN (k) MNIST FastKAN (l) MNIST FasterKAN

(m) CIFAR-10 EfficientKAN (n) CIFAR-10 FastKAN (o) CIFAR-10 FasterKAN (p) CIFAR-100 EfficientKAN

(q) CIFAR-100 FastKAN (r) CIFAR-100 FasterKAN

图 3.4 水印任务上的分类准确率：(a)-(i)剪枝攻击后性能评估；(j)-(r)微调攻击后性能评估

从图 3.3 和图 3.4 中的(a)-(i)可以推断出，对于剪枝攻击，KAN 模型的分类性能随着剪枝率的增加而逐渐下降。然而，即使剪枝率达到 50%，KAN 模型在水印任务上的分类准确率仍然能够达到 100%，这表明所提出的方法对剪枝攻击具有很强的抵抗力。

从图 3.3 和图 3.4 中的(j)-(r)可以推断出，对于微调攻击，KAN 模型在使用微调时嵌入的水印嵌入策略后，其在水印任务上的分类性能表现最差。尽管在某些情况下，使用训练时嵌入的水印嵌入策略并且使用不同比例的验证集对 KAN 模型进行微调后，KAN 模型在水印任务上的表现仍然能够维持在相对较高的水平，但这种方法缺乏普遍性。相比之下，使用本章提出的方法训练的 KAN 模型在遭受微调攻击后，在水印任务上仍能保持超过 98.67%的平均分类准确率，同时保持了 KAN 模型在原始任务上的性能，从而证明了所提出的方法对微调攻击的强大抵抗力。

3.3.4 对比实验

本节我们将所提出的方法与文献[26]和文献[27]的方法进行了对比，分别在 KAN 模型的原始任务、水印任务上比较分类准确率，并验证 KAN 模型水印的鲁棒性。其中，所采用的 KAN 模型为 FastKAN，所选数据集为 CIFAR-10。

表 3.5 展示了不同方法下水印嵌入后 KAN 模型在原始任务和水印任务上的分类准确率。表 3.6 展示了不同方法下使用不同剪枝率对 KAN 模型在原始任务和水印任务上的分类准确率的影响。表 3.7 展示了不同方法下使用不同微调比例对 KAN 模型在原始任务和水印任务上的分类准确率的影响。从表 3.5 可以推断出，文献[26]和文献[27]的方法以及本章提出的方法均能够实现 100%的水印任务的分类准确率，但是在原始任务的分类准确率上本章方法的表现最优，相较于文献[26]和文献[27]的方法分别提升了 2.08%和 0.96%，表明本章方法在嵌入水印的同时对 KAN 模型原始任务的性能影响更小。从表 3.6 可以推断出，随着剪枝率的逐步增加，三种方法在原始任务上的分类准确率均呈现下降趋势，但是本章方法展现出更强的鲁棒性。具体而言，当剪枝率达到 50%时，文献[26]和文献[27]的水印任务分类准确率已经分别降至 88%和 99%，而本章提出的方法仍然能够保持在 100%。即使在 90%的高强度剪枝攻击下，本章方法的水印任务分类准确率仍然达到了 73%，远高于文献[26]和文献[27]的 5%和 35%，表明本章方法在高强度剪枝攻击下依然能够有效保护水印信息。从表 3.7 可以推断出，随着微调比例的逐步增加，三种方法在原始任务上的分类准确率均出现一定程度的下降，虽然整体变化幅度较小，但是本章方法的分类性能总体上是要高于文献[26]和文献

[27]。在水印鲁棒性方面，三种方法的表现各不相同。具体而言，文献[26]和文献[27]在微调攻击下，水印任务的分类准确率均出现了不同程度的下降，其中以文献[26]的下降幅度最大，平均降低了约为 18%，而文献[27]的平均下降幅度则约为 4%，并且文献[26]在微调比例达到 20%时，在水印任务上的分类准确率达到最低值，而文献[27]在微调比例达到 70%及以上时，在水印任务上的分类准确率才达到最低值。相比于文献[26]和文献[27]，本章方法在各微调比例下均能够保持 100%的水印任务分类准确率，表明本章提出的方法对微调攻击具有更强的抵抗能力，水印稳定性更佳。综上所述，本章方法要优于文献[26]和文献[27]的方法，在保持模型准确率相当的同时，对水印去除攻击具有更强的抵抗力，这证明了所提出的基于频域扰动和交替训练的鲁棒黑盒KAN模型水印方法在提高KAN模型安全性方面的有效性。

表 3.5 不同方法下 KAN 模型在原始任务和水印任务上的分类准确率

方法	Acc_o (%)	Acc_w (%)
文献[26]	52.01	100
文献[27]	53.13	100
本章方法	54.09	100

表 3.6 不同剪枝率下各方法的原始任务和水印任务的分类准确率

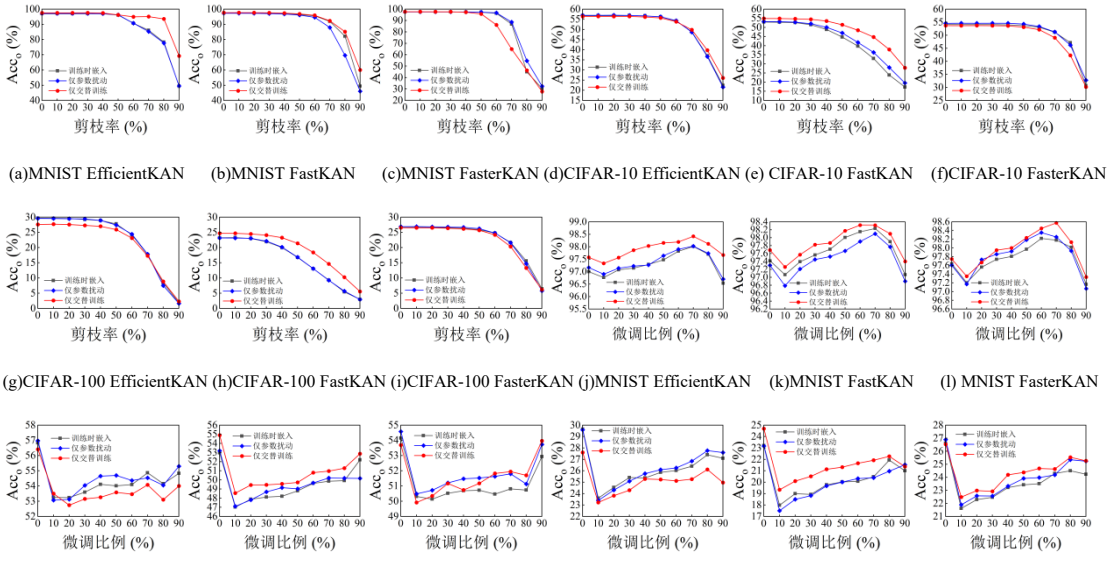
剪枝率(%)	文献[26]		文献[27]		本章方法	
	Acc_o (%)	Acc_w (%)	Acc_o (%)	Acc_w (%)	Acc_o (%)	Acc_w (%)
10	51.91	100	53.05	100	54.02	100
20	51.71	100	52.64	100	54.08	100
30	51.27	100	51.74	100	53.61	100
40	48.43	95.00	48.58	100	52.98	100
50	43.59	88.00	44.82	99.00	51.78	100
60	37.27	78.00	40.42	96.00	47.37	100
70	32.57	52.00	33.44	89.00	43.42	100
80	26.65	16.00	28.79	70.00	37.95	100
90	18.87	5.00	20.35	35.00	28.34	73.00

表 3.7 不同微调比例下各方法的原始任务和水印任务的分类准确率

微调比例(%)	文献[26]		文献[27]		本章方法	
	Acc_o (%)	Acc_w (%)	Acc_o (%)	Acc_w (%)	Acc_o (%)	Acc_w (%)
10	45.97	97.00	46.71	96.00	50.10	100
20	46.35	73.00	47.28	96.00	49.14	100
30	47.91	84.00	47.81	97.00	50.57	100
40	48.72	81.00	48.98	97.00	50.07	100
50	48.16	82.00	49.18	97.00	49.78	100
60	48.38	82.00	50.15	97.00	50.63	100
70	49.07	80.00	49.63	94.00	51.40	100
80	47.85	78.00	49.80	94.00	50.85	100
90	50.30	80.00	52.10	95.00	53.50	100

3.3.5 消融实验

为了全面评估所提出的方法的性能，我们仅利用参数扰动和仅使用交替训练进行了消融实验。鉴于采用微调时嵌入策略训练的 KAN 模型无法有效抵御剪枝攻击和微调攻击，我们比较了采用训练时嵌入、仅利用参数扰动以及仅使用交替训练的 KAN 模型在剪枝攻击和微调攻击下的性能，结果如图 3.5 和图 3.6 所示。



(a)MNIST EfficientKAN (b)MNIST FastKAN (c)MNIST FasterKAN (d)CIFAR-10 EfficientKAN (e) CIFAR-10 FastKAN (f)CIFAR-10 FasterKAN

(g)CIFAR-100 EfficientKAN (h)CIFAR-100 FastKAN (i)CIFAR-100 FasterKAN (j)MNIST EfficientKAN (k)MNIST FastKAN (l) MNIST FasterKAN

(m)CIFAR-10 EfficientKAN(n)CIFAR-10 FastKAN(o)CIFAR-10 FasterKAN (p)CIFAR-100 EfficientKAN(q)CIFAR-100 FastKAN(r)CIFAR-100 FasterKAN

图 3.5 原始任务上的分类准确率：(a)-(i)剪枝攻击后性能评估；(j)-(r)微调攻击后性能评估

(a) MNIST EfficientKAN (b) MNIST FastKAN (c) MNIST FasterKAN (d) CIFAR-10 EfficientKAN

(e) CIFAR-10 FastKAN (f) CIFAR-10 FasterKAN (g) CIFAR-100 EfficientKAN (h) CIFAR-100 FastKAN

(i) CIFAR-100 FasterKAN (j) MNIST EfficientKAN (k) MNIST FastKAN (l) MNIST FasterKAN

(m) CIFAR-10 EfficientKAN (n) CIFAR-10 FastKAN (o) CIFAR-10 FasterKAN (p) CIFAR-100 EfficientKAN

(q) CIFAR-100 FastKAN (r) CIFAR-100 FasterKAN

图 3.6 水印任务上的分类准确率：(a)-(i)剪枝攻击后性能评估；(j)-(r)微调攻击后性能评估

从图 3.5 和图 3.6 的(a)-(i)可以推断出，对于剪枝攻击，在相同的剪枝率条件下，无论是采用仅利用参数扰动还是仅使用交替训练，都存在一种方法在水印任务上的分类准确率优于采用训练时嵌入的方法，同时保持了 KAN 模型在原始任务上的性能，这表明参数扰动和交替训练在抵御剪枝攻击方面具有显著的协同效

应和互补性。从图 3.5 和图 3.6 的(j)-(r)可以推断出，对于微调攻击，在相同的微调比例条件下，仅利用参数扰动或仅使用交替训练的 KAN 模型在水印任务上的分类准确率始终高于采用训练时嵌入的模型，同时保持了 KAN 模型在原始任务上的性能，这表明参数扰动和交替训练在提高水印任务的分类准确率方面表现出令人满意的性能，并有效增强了 KAN 模型对微调攻击的鲁棒性。

此外，我们还评估了在不同a_l值（$a_l \in \{0.001, 0.01, 0.1\}$）下，KAN 模型在 CIFAR-10 数据集上的平均性能，结果如表 3.8 所示。从表 3.8 中可以看出，将α_l从 0.001 增加到 0.1 后能够使模型在剪枝攻击下的水印任务的分类准确率从 96.11%提高到 99.89%，但是却导致模型在原始任务上的分类准确率出现了一定程度的下降（剪枝后下降 0.69%，微调后下降 0.33%）。因此，将α_l设置为 0.01 既能保证足够的水印鲁棒性，又能兼顾原始任务的性能。

表 3.8 CIFAR-10 数据集上$\alpha_l \in \{0.001, 0.01, 0.1\}$时 KAN 模型的平均性能

α_l	初始值		70%剪枝率		70%微调比例	
	Acc_o (%)	Acc_w (%)	Acc_o (%)	Acc_w (%)	Acc_o (%)	Acc_w (%)
0.001	54.78	100	46.75	96.11	53.78	100
0.01	54.95	100	46.04	98.44	53.79	100
0.1	54.82	100	46.06	99.89	53.45	100

3.3.6 计算成本分析

对所提出的方法的计算成本分析是通过系统地评估 KAN 模型在每个周期中无水印模型训练时间、含水印模型训练时间以及最终水印验证消耗的时间，结果如表 3.9 所示。实验结果表明，水印嵌入会引入额外的开销，其大小取决于数据集和模型的配置。在 MNIST 数据集上使用 FasterKAN 模型时，额外时间成本仅约为无水印模型训练时间的 32%，这表明更简单的数据集和更高效的 KAN 模型可以有效降低相对成本。同时，对于这三个数据集上的所有 KAN 模型，水印验证所使用的时间均在 1-3 毫秒以内，这表明一旦训练完成，后门检测几乎不会增加额外的计算负担。总体而言，所提出的方法不仅能够在不影响 KAN 模型在原始任务上的性能的情况下嵌入鲁棒的水印，而且在实践应用中引入了可接受的额外成本，凸显了其在模型所有权验证和保护 KAN 模型方面的强大潜力。

表 3.9 不同 KAN 模型训练与验证阶段计算时间比较

数据集	模型	无水印模型训练时间(s)	含水印模型训练时间(s)	水印验证时间(ms)
MNIST	EfficientKAN	6.59	9.80	1.63
	FastKAN	6.22	8.58	1.31
	FasterKAN	5.93	7.85	1.07
CIFAR-10	EfficientKAN	7.05	13.74	2.79
	FastKAN	6.23	10.54	1.41
	FasterKAN	5.79	9.55	1.40
CIFAR-100	EfficientKAN	7.66	13.82	2.85
	FastKAN	6.51	10.81	1.60
	FasterKAN	6.10	9.64	1.52

3.4 本章小结

在黑盒条件下保护深度神经网络模型的知识产权，尤其是对于新兴的 KAN 模型，是一项极具挑战性和至关重要的任务。现有的黑盒水印方法虽然能在一定程度上验证模型所有权，但尚未深入探索如何协同使用训练集和触发集来训练目标模型。此外，这些方法对水印去除攻击的鲁棒性仍有待提高。为此，本章提出了一种基于频域扰动和交替训练的鲁棒黑盒 KAN 模型水印方法。该方法采用交替训练策略，交替使用原始训练集和触发集进行模型训练，同时引入模型参数扰动损失来模拟水印去除攻击，有效增强了水印的鲁棒性。实验结果表明，所提出的方法在不显著影响 KAN 模型性能的前提下，即使面对 50%的剪枝率，也可以在水印任务上实现 100%的分类准确率。此外，在三个不同的触发集T_{DWT}、T_{DCT}和T_{FFT}上，经微调攻击后，KAN 模型在水印任务上的平均性能仍然能够保持在 98.67%以上，验证了所提出的方法的优秀性能。

第四章　基于激活扰动的 KAN 模型水印

4.1 引言

随着新兴神经网络 KAN 模型的诞生，其独特的架构和特性不断拓展着自身的应用范围和商业价值，与此同时，模型盗用与模型滥用等问题也日益严重。因此，KAN 模型的知识产权保护问题愈发凸显，对其水印保护技术的设计提出了更高的要求。在上一章中，提出了一种基于频域扰动和交替训练的鲁棒黑盒 KAN 模型水印，该方法将水印信息嵌入图像频域以构建触发集，然后通过交替使用原始任务训练集和触发集训练目标 KAN 模型，同时通过施加参数扰动来模拟水印去除攻击，使得水印在常见的水印去除攻击下依然能够成功验证，有效地保证了水印的鲁棒性。然而，由于 KAN 模型的参数主要分布在各边的可学习激活函数上，其水印嵌入过程会直接体现为激活函数形态的改变，使得水印 KAN 模型和干净 KAN 模型在相应边上的激活函数产生明显的形态差异，从而导致水印的嵌入缺乏隐蔽性。

为了应对上述挑战，本章提出一种基于激活扰动的 KAN 模型水印方案，通过对训练好的干净 KAN 模型的第一层激活函数进行轻微扰动来完成水印信息的嵌入，从而提高水印在激活函数上的形态不可感知性。此外，为了提高水印鲁棒性，本章进一步引入参数对抗训练，通过对 KAN 模型的激活函数施加一定程度的随机噪声来模仿 KAN 模型遭受水印去除攻击的情况，使得水印验证网络仍能成功提取水印。因此，本章提出了一种适用于 KAN 模型的水印方案，并在后续实验中验证了所提方案在不同分类任务场景下水印的形态不可感知性和鲁棒性。

4.2 基于激活扰动的 KAN 模型水印方案

4.2.1 总体框架

本章提出了一种基于激活扰动的 KAN 模型水印方案，旨在保证水印鲁棒性

的同时能够提高水印激活函数的形态不可感知性。该方案的总体框架如图 4.1 所示，不同于需要使模型从头开始一起训练原始任务和水印任务，所提出的水印框架只需要让 KAN 模型在原始任务训练好后迭代训练较少轮次即可完成水印的嵌入，嵌入过程灵活，所需计算资源更少，具有更好的实际应用价值。所提出的框架主要由三个模块组成，即：原始任务训练后的干净 KAN 模型，水印嵌入网络 E 和水印提取网络 R。KAN 模型被广泛应用于回归任务、时序预测、图像分类等任务中，这里干净 KAN 模型的任务是执行与图像分类相关的任务。在训练阶段 A，KAN 模型将完成图像分类任务的学习从而得到干净 KAN 模型。在训练阶段 B，原始任务训练后的干净 KAN 模型，水印嵌入网络 E 和水印提取网络 R 将一起被训练。具体而言，水印嵌入网络 E 将水印信息嵌入到接收到的干净 KAN 模型第一层的激活函数上从而得到水印 KAN 模型。在进行模型水印提取之前，水印 KAN 模型可能会遭受各种模型水印去除攻击。为了确保嵌入的水印信息能抵抗常见的模型水印去除攻击，一种普遍的方法是在模型训练期间引入额外的对抗训练来模拟实际的攻击情况，因此，本章在水印嵌入网络 E 和水印提取网络 R 之间增加了一个模拟攻击层，用以模拟水印 KAN 模型在遭受模型水印去除攻击后的情形，要求水印提取网络 R 仍能够从水印 KAN 模型中成功提取出水印信息。

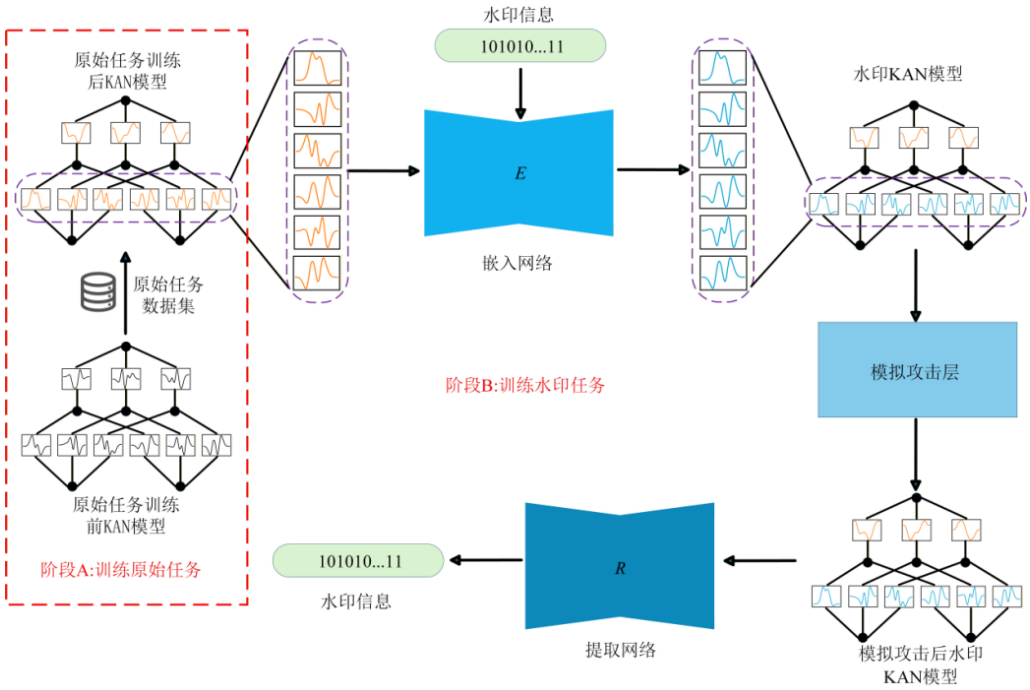


图 4.1 基于激活函数扰动的 KAN 模型水印框架

4.2.2 损失函数

本章所提出的 KAN 模型水印框架包括两个训练阶段。阶段 A 是训练 KAN 模型来完成原始任务，阶段 B 则是训练已经在原始任务上训练完毕的不含水印的干净 KAN 模型，水印嵌入网络 E 和水印提取网络 R 来完成水印任务。阶段 A 和阶段 B 相互独立。

阶段 A：KAN 模型负责完成图像分类相关的任务，在训练过程中，模型的损失函数使用交叉熵损失函数。

阶段 B：为了实现模型水印任务，水印嵌入网络 E 和水印提取网络 R 将会和阶段 A 中已经在原始任务上训练完毕的不含水印的干净 KAN 模型一起被训练。首先，本章希望水印嵌入网络 E 在将水印信息嵌入到接收到的干净 KAN 模型第一层的激活函数上后与将水印信息嵌入前的 KAN 模型第一层的激活函数的形态要尽可能相似，故构造损失函数 $\mathcal{L}_E$ 来约束嵌入水印后的 KAN 模型第一层激活函数的响应特征和嵌入水印前的 KAN 模型第一层激活函数的响应特征的偏离程度，同时通过方差正则项防止响应特征坍缩，$\mathcal{L}_E$ 被定义为公式(4.1)：

$$\mathcal{L}_E = \frac{1}{B \cdot d}\sum_{i=1}^{B}\sum_{j=1}^{d}(f_{ij}^{\text{post}} - f_{ij}^{\text{pre}})^2 - \alpha \cdot \frac{1}{d}\sum_{j=1}^{d}\mathbb{E}_i\left[(f_{ij}^{\text{post}} - \mathbb{E}_i[f_{ij}^{\text{post}}])^2\right] \quad (4.1)$$

其中，$\mathbb{E}_i[\cdot] = 1/B\sum_{i=1}^{B}(\cdot)$，$B$ 为批量大小，d 为特征维度，f_{ij}^{pre} 是嵌入水印信息前的 KAN 模型第一层激活函数的响应特征，f_{ij}^{post} 是嵌入水印信息后的 KAN 模型第一层激活函数的响应特征，α 被设置为 0.001。同时为了约束嵌入水印信息后 KAN 模型第一层的参数与嵌入水印信息前 KAN 模型第一层的参数的偏离程度，构造权重守恒损失 $\mathcal{L}_W$，其被定义为公式(4.2)：

$$\mathcal{L}_W = \sum_{p\in\theta_0} \| p - p^{(0)} \|_2^2 \quad (4.2)$$

其中 θ_0 表示为 KAN 模型第一层所有可训练的模型参数，p 表示为嵌入水印信息后的第一层 KAN 模型参数，$p^{(0)}$ 表示为未嵌入水印信息的第一层 KAN 模型参数。对于水印提取网络R，其约束条件是从含有水印信息的 KAN 模型第一层的激活函数中能够正确提取出嵌入的水印信息。此约束通过优化水印提取损失 $\mathcal{L}_R$ 来实现，$\mathcal{L}_R$ 用于衡量水印提取网络R从嵌入水印信息后的 KAN 模型第一层激活

函数的响应特征中准确恢复预设的L位水印信息比特序列$w \in \{0,1\}^L$的能力，其被定义为公式(4.3)：

$$\mathcal{L}_R = \frac{1}{B \cdot L} \sum_{i=1}^{B} \sum_{k=1}^{L} (R(f_i^{\text{post}})_k - w_k)^2 \tag{4.3}$$

同时为了确保嵌入水印信息前的干净 KAN 模型的第一层激活函数的响应特征不会被误检出预设的水印信息，而是只能从中提取出噪声信息，因此构造损失函数$\mathcal{L}_R^{clean}$，其被定义为公式(4.4)：

$$\mathcal{L}_R^{\text{clean}} = \frac{1}{B \cdot L} \sum_{i=1}^{B} \sum_{k=1}^{L} (R(f_i^{\text{clean}})_k - \tilde{w}_k)^2 \tag{4.4}$$

其中$\tilde{w} \in \{0,1\}^L$为随机生成的比特序列，该损失迫使干净 KAN 模型第一层激活函数响应特征的提取结果为随机比特序列。为了增强模型水印对水印去除攻击的鲁棒性，通过添加轻微高斯噪声的方式来对含有水印信息的 KAN 模型第一层的激活函数进行轻微的扰动，要求提取网络 R 仍然能够从嵌入水印信息后的 KAN 模型第一层激活函数的响应特征中准确恢复预设的水印信息比特序列，因此构造损失函数 $\mathcal{L}_R^{\text{noise}}$，其被定义为公式(4.5)：

$$\mathcal{L}_R^{\text{noise}} = \frac{1}{B \cdot L} \sum_{i=1}^{B} \sum_{k=1}^{L} (R(f_i^{\text{post}} + \beta \cdot \epsilon_i)_k - w_k)^2 \tag{4.5}$$

其中$\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$，$\epsilon_i$是从均值为零、方差为$\sigma_i^2$的高斯分布中采样的噪声项，$\sigma_i^2$是第$i$批次样本经 KAN 模型第一层输出 f_i^{post} 的标准差，β用于调整噪声项的大小，被设置为 0.01。此外，为了在水印嵌入过程中保持 KAN 模型的原始分类性能，防止过拟合水印模式而遗忘原始任务知识，构造任务保持损失函数 $\mathcal{L}_P$，其被定义为公式(4.6)：

$$\mathcal{L}_P = L_{CE}(KAN(X), Y) \tag{4.6}$$

其中X为输入样本，Y为对应的标签，L_{CE}为交叉熵损失函数。最终，原始任务训练后的干净 KAN 模型，水印嵌入网络 E 和水印提取网络 R 的总损失函数可以表述为公式(4.7)：

$$\mathcal{L}_{\text{total}} = \lambda_E \mathcal{L}_E + \lambda_R \mathcal{L}_R + \lambda_n \mathcal{L}_R^{\text{noise}} + \lambda_c \mathcal{L}_R^{\text{clean}} + \lambda_W \mathcal{L}_W + \lambda_P \mathcal{L}_P \tag{4.7}$$

其中 λ_E， λ_R， λ_n， λ_c， λ_W 和 λ_P 都是可调超参数。

4.3 实验结果与分析

4.3.1 实验设置

数据集和模型：为了验证所提出方案的有效性，本节评估了它在 Fashion-MNIST[83]，MNIST[81]，CIFAR-10[82]和 CIFAR-100[82]这四个数据集上的表现。KAN 模型选用 EfficientKAN，FastKAN 和 FasterKAN。水印嵌入网络 E 和水印提取网络 R 都使用一个三层的多层感知机。

超参数设定：对于可调超参数，λ_E 被设置为 1，λ_R 被设置为 3，λ_n 被设置为 0.1，λ_c 被设置为 1，λ_W 被设置为 0.03，λ_P 被设置为 0.5。

训练设置：在使用本章提出的方法训练 KAN 模型时，我们采用两阶段训练策略。第一阶段为干净模型训练阶段：我们使用 AdamW 优化器对干净 KAN 模型进行参数优化，初始学习率设置为 0.001，权重衰减设置为 0.0001。学习率通过指数调度器进行调整，其中 gamma 设置为 0.8，调度器从第 5 个周期开始生效。我们总共对干净 KAN 模型进行 50 个周期的训练，每个周期在训练集上优化模型参数，并在验证集上评估分类性能。第二阶段为水印嵌入训练阶段：冻结 KAN 模型除第一层外的所有模型参数，仅优化模型第一层参数、水印嵌入网络 E和水印提取网络R。此阶段使用单独的 AdamW 优化器，初始学习率设置为 0.0001，权重衰减设置为 0.0001，学习率通过指数调度器进行调整，其中 gamma 设置为 0.8，调度器每个周期都生效。总共进行 20 个周期的水印嵌入训练。所有实验均在配备 NVIDIA GeForce RTX 4060 GPU 和 Intel(R) Core(TM) i7-14650HX CPU 以及 16.0 GB 内存的计算机环境中进行。

4.3.2 保真度分析

任务保真度要求 KAN 模型在嵌入水印后，其原始任务的分类准确率不能显著降低。表 4.1 详细列出了水印嵌入前后 KAN 模型的性能评估。从表 4.1 中可以看出，不同 KAN 模型在各数据集上的性能表现存在显著差异，这既反映了不

同激活函数在学习能力上的区别，也体现了数据集复杂度对 KAN 模型性能的影响。在 Fashion-MNIST 数据集上，EfficientKAN 模型、FastKAN 模型和 FasterKAN 模型的原始任务的分类准确率分别为 88.79%、89.67%和 89.36%，在模型嵌入水印后分别变为 88.71%、89.57%和 89.42%，下降幅度均控制在 0.12%以内，其中 FasterKAN 模型甚至略有提升。在 MNIST 数据集上，三种模型的原始准确率均能够保持在 97%以上的高水平，嵌入水印后的模型分类性能下降幅度同样微乎其微，EfficientKAN 模型从 97.40%降至 97.29%，FastKAN 模型则从 97.53%降至 97.44%，FasterKAN 模型则从 97.89%降至 97.87%。对于更为复杂的 CIFAR-10 数据集，EfficientKAN 模型和 FastKAN 模型的分类准确率下降了约为 0.02%和 0.08%，而 FasterKAN 模型则从 53.15%提升至 53.29%。在最具挑战性的 CIFAR-100 数据集上，EfficientKAN 模型和 FasterKAN 模型基本保持稳定，FastKAN 模型虽有较明显下降（从 24.56%降至 23.69%），但考虑到该数据集本身的高难度和模型本身的较低基准，这一波动在可接受范围内。因此，本章提出的方法能够有效降低 KAN 模型在原始任务上分类准确率的衰减。特别是在 Fashion-MNIST 和 MNIST 这类相对简单的数据集上，水印 KAN 模型的性能保持尤为稳定；即使在 CIFAR-10 和 CIFAR-100 这类复杂数据集上，水印 KAN 模型也能够维持原有的性能水平，部分情况下甚至有所提升。这表明本章提出的方法充分利用了 KAN 模型的冗余性，在不损害原始任务分类能力的前提下完成水印嵌入，在实际应用中具有巨大的潜力。

此外，水印保真度要求嵌入 KAN 模型中的水印能够被准确重构，从而确保水印检测的可靠性。在本章中，我们的目标是要在水印 KAN 模型中能够准确检测出嵌入的水印信息。从表 4.1 中 KAN 模型嵌入水印后在水印任务上的分类准确率一列可以看出，实验所采用的三类 KAN 模型，在 Fashion-MNIST、MNIST、CIFAR-10 和 CIFAR-100 这四个数据集上，水印任务的分类准确率均能够达到 100%。这一结果充分证明了本章提出的方法具有卓越的通用性和优越性，它不仅能够适应不同架构的 KAN 模型，还能在不同复杂度的数据集上保持稳定的水印嵌入效果。

表 4.1 水印嵌入前后 KAN 模型的性能

数据集	KAN 模型	嵌入水印前	嵌入水印后	
		Acc_o (%)	Acc_o (%)	Acc_w (%)
Fashion-MNIST	EfficientKAN	88.79	88.71	100
	FastKAN	89.67	89.57	100
	FasterKAN	89.36	89.42	100
MNIST	EfficientKAN	97.40	97.29	100
	FastKAN	97.53	97.44	100
	FasterKAN	97.89	97.87	100
CIFAR-10	EfficientKAN	55.58	55.56	100
	FastKAN	54.53	54.45	100
	FasterKAN	53.15	53.29	100
CIFAR-100	EfficientKAN	27.44	27.40	100
	FastKAN	24.56	23.69	100
	FasterKAN	25.63	25.67	100

4.3.3 水印形态不可感知性分析

为了分析水印信息嵌入前后对 KAN 模型第一层激活函数的形态差异，首先在 Fashion-MNIST，MNIST，CIFAR-10 和 CIFAR-100 这四个数据集上可视化 KAN 模型第零层中第100个神经元连接到第一层第50个神经元的激活函数 $\phi_{0,50,100}$，结果如图 4.2 所示。从图 4.2 可以观察到，水印嵌入前后 KAN 模型的激活函数曲线整体形态能够保持高度相似，在视觉上几乎看不出激活函数形态的差别，表明水印嵌入并未破坏 KAN 模型激活函数的基本非线性特征，具有良好的形态不可感知性。

进一步地，为定量评估水印嵌入对 KAN 模型激活函数形态的影响，本节通过平均绝对误差(Mean Absolute Error, MAE)度量干净 KAN 模型与水印 KAN 模型在相同输入样本处激活值的平均数值偏差并且引入皮尔逊相关系数(Pearson Correlation Coefficient, PCC)评估上述这两组激活值序列的线性相关性与形态保真度，其中，MAE 值越小表明水印对激活函数数值精度的影响越轻微而 PCC 值

越接近数值 1 表明水印嵌入后激活函数的形态保持越完整。具体而言，对于将 KAN 模型第零层的第100个神经元连接到第一层第50个神经元的激活函数$\phi_{0,50,100}$，在输入定义域$\mathcal{X}=[-1,1]$内均匀采样$n=300$个数据点 $\{x_k\}_{k=1}^{n}$，然后分别计算 MAE 值和 PCC 值，其形式分别如公式(4.8)和(4.9)所示：

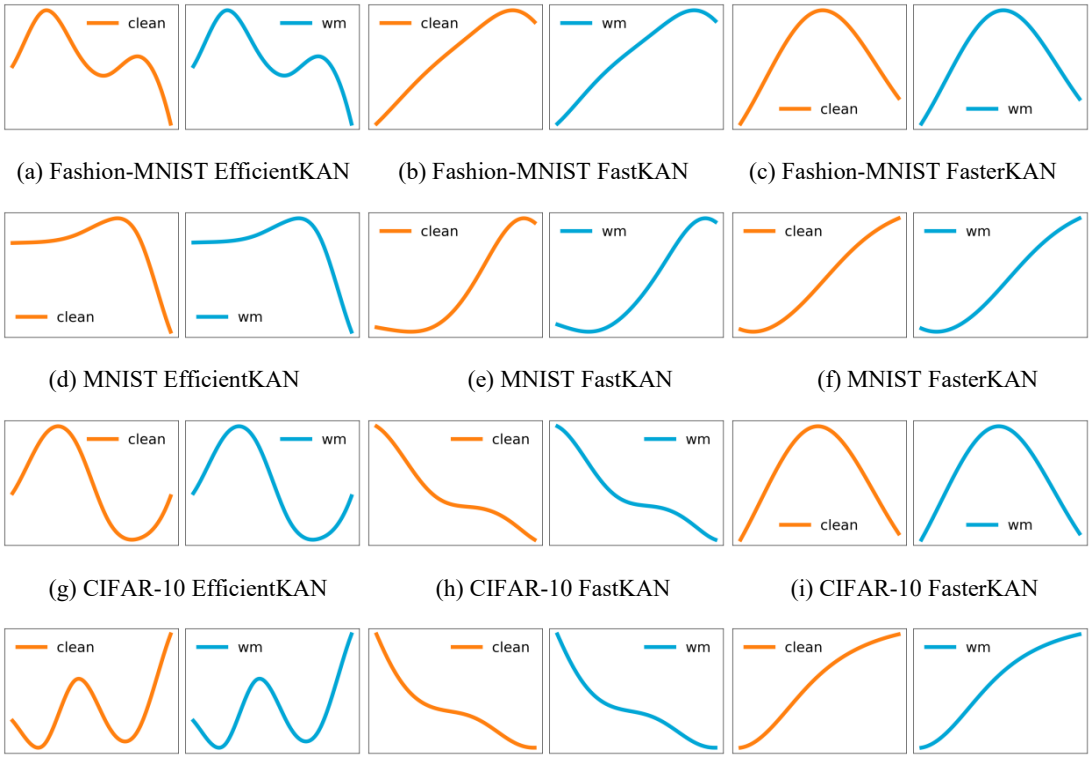
$$MAE=\frac{1}{n}\sum_{k=1}^{n}\left|\phi_{0,50,100}^{\text{clean}}(x_k)-\phi_{0,50,100}^{\text{wm}}(x_k)\right| \tag{4.8}$$

$$PCC=\frac{\sum_{k=1}^{n}(\phi_{0,50,100}^{\text{clean}}(x_k)-\overline{\mu}_c)(\phi_{0,50,100}^{\text{wm}}(x_k)-\overline{\mu}_w)}{\sqrt{\sum_{k=1}^{n}(\phi_{0,50,100}^{\text{clean}}(x_k)-\overline{\mu}_c)^2}\sqrt{\sum_{k=1}^{n}(\phi_{0,50,100}^{\text{wm}}(x_k)-\overline{\mu}_w)^2}} \tag{4.9}$$

其中$\overline{\mu}_c=1/n\sum_{k=1}^{n}\phi_{0,50,100}^{\text{clean}}(x_k)$和$\overline{\mu}_w=1/n\sum_{k=1}^{n}\phi_{0,50,100}^{\text{wm}}(x_k)$分别表示为干净 KAN 模型第一层激活函数响应值的均值和水印 KAN 模型第一层激活函数响应值的均值，结果如表 4.2 所示。从表 4.2 的数据可以看出，三类 KAN 模型在四个数据集上均能够展现出较高的激活函数保真度。在 MAE 指标方面，EfficientKAN 模型表现最优，其在 CIFAR-10 数据集和 CIFAR-100 数据集上的 MAE 值低至10^{-7}量级，在 Fashion-MNIST 数据集和 MNIST 数据集上也能够保持在10^{-5}量级，表明水印的嵌入对其激活函数数值的影响微乎其微。相比之下，FastKAN 模型则波动较大，其 MAE 范围在 0.0022~0.0084 之间，在 Fashion-MNIST 数据集上更是达到最高值，相比 EfficientKAN 模型高出约 460 倍，说明 FastKAN 模型的激活函数对水印嵌入更为敏感。FasterKAN 模型则表现居中，其 MAE 范围在 0.0002~0.0038 之间，在 MNIST 数据集上表现相对较差，但在 CIFAR-10 数据集上已经能够较为接近 EfficientKAN 模型的水平。在 PCC 指标方面，所有水印 KAN 模型的 PCC 值均大于 0.998，绝大多数接近 0.9999 甚至 1.0000。其中，EfficientKAN 模型表现最优，其在 CIFAR-10 数据集上能够达到 1.00000000，在 CIFAR-100 数据集上能够达到 0.99999946，在 Fashion-MNIST 数据集上也依然能够达到 0.99999797，表明水印嵌入后其激活函数形态几乎能够完全保持。FastKAN 模型的形态保真度则相对较低，其在 MNIST 数据集上降至 0.99873495，为全部数据中的最低值，但仍然能够保持较高的线性相关性。FasterKAN 模型则表现最为稳定，各数据集上 PCC 值均保持在 0.9995 以上，尤其在 CIFAR-10 数据集上能够

达到 0.99999976。综合对比来看，EfficientKAN 模型在所有四个数据集上均同时取得 MAE 指标和 PCC 指标的最优值，展现出最强的激活函数鲁棒性。此外，这三类 KAN 模型在 CIFAR 系列数据集上的 MAE 值显著低于在 MNIST 系列数据集上的值，这可能与 CIFAR 系列数据集的复杂特征分布使 KAN 模型学习到更加鲁棒的激活模式有关。从 KAN 模型的结构角度分析，FastKAN 模型使用径向基函数作为激活函数，水印嵌入可能导致其可学习参数发生相对明显的偏移，从而表现出较高的 MAE 值，说明其激活函数具有一定的敏感性。而作为 FastKAN 模型的优化版本，FasterKAN 模型在保持较高计算效率的同时，通过改进的激活函数和模型训练机制，将 MAE 值降低了约 50%~80%。

综上所述，本章所提出的水印嵌入方法能够向 KAN 模型第一层激活函数中引入较小的扰动数值，并且水印嵌入后激活函数的几何形态、单调性、曲率特征等关键属性均能够得到较为完整的保持，具有较高的激活函数保真度和良好的形态不可感知性。



(a) Fashion-MNIST EfficientKAN (b) Fashion-MNIST FastKAN (c) Fashion-MNIST FasterKAN

(d) MNIST EfficientKAN (e) MNIST FastKAN (f) MNIST FasterKAN

(g) CIFAR-10 EfficientKAN (h) CIFAR-10 FastKAN (i) CIFAR-10 FasterKAN

(j) CIFAR-100 EfficientKAN (k) CIFAR-100 FastKAN (l) CIFAR-100 FasterKAN

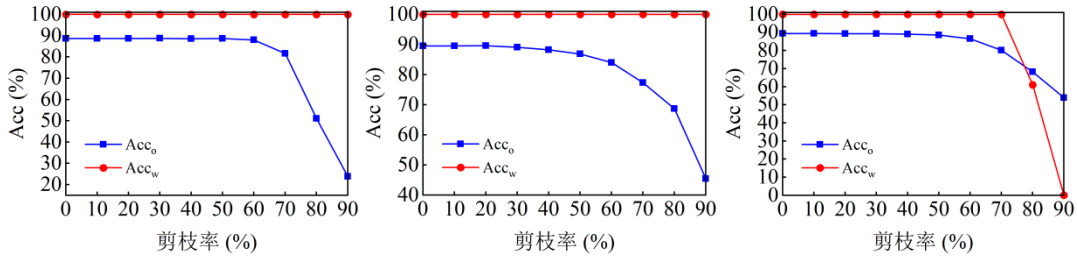
图 4.2 水印嵌入前后 KAN 模型激活函数$\phi_{0,50,100}$的可视化对比

表 4.2 水印嵌入前后 KAN 模型激活函数$\phi_{0,50,100}$的误差与相关性分析

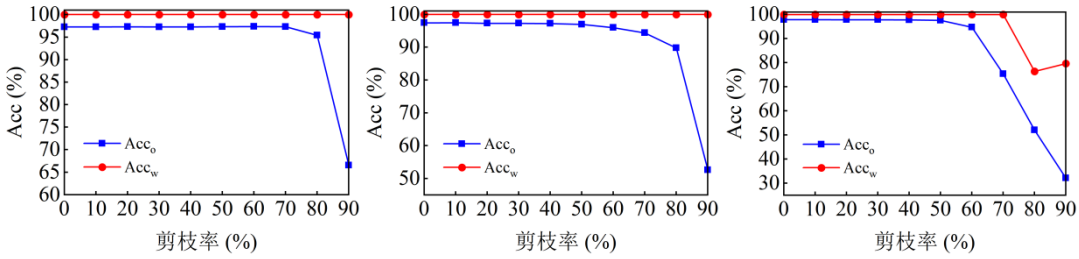
数据集	KAN 模型	MAE	PCC
Fashion-MNIST	EfficientKAN	0.00001821	0.99999797
	FastKAN	0.00836766	0.99984902
	FasterKAN	0.00118598	0.99995083
MNIST	EfficientKAN	0.00007995	0.99996412
	FastKAN	0.00820772	0.99873495
	FasterKAN	0.00376184	0.99959558
CIFAR-10	EfficientKAN	0.00000054	1.00000000
	FastKAN	0.00220979	0.99972630
	FasterKAN	0.00020246	0.99999976
CIFAR-100	EfficientKAN	0.00000023	0.99999946
	FastKAN	0.00393123	0.99997222
	FasterKAN	0.00157111	0.99997270

4.3.4 鲁棒性分析

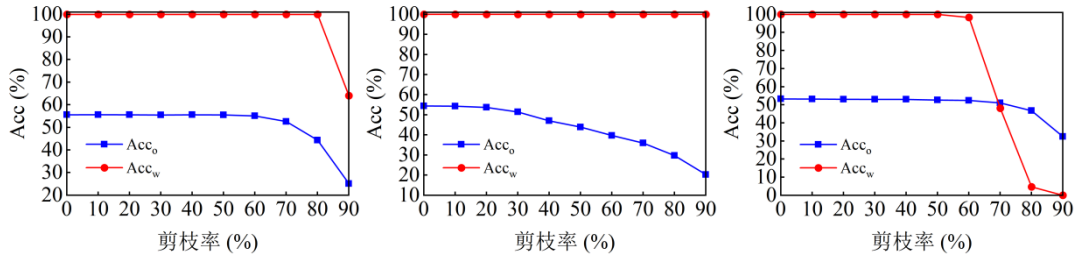
为了评估本章所提出的 KAN 模型水印的鲁棒性，我们比较了四种常见的应用于模型的水印去除攻击，它们分别是剪枝攻击、微调攻击、剪枝-微调攻击和覆盖攻击。对于剪枝攻击，我们对水印 KAN 模型应用了L_1范数剪枝策略，其中剪枝率表示被剪枝参数的百分比。对于微调攻击，我们使用不同比例的验证集对水印 KAN 模型进行微调，其中微调比例表示用于模型微调的验证集的比例。对于剪枝-微调攻击，我们先对水印 KAN 模型进行 50%剪枝率的剪枝攻击，再对水印 KAN 模型进行 50%微调比例的微调攻击。对于覆盖攻击，我们则是向已经嵌入水印信息的 KAN 模型内使用同样的水印嵌入方式再次植入新的水印信息。同样地，我们也定义了Acc_o和Acc_w，其中Acc_o表示 KAN 模型在原始任务上的分类准确率，Acc_w表示 KAN 模型在水印任务上的分类准确率。 图 4.3 展示了剪枝攻击后水印 KAN 模型在原始任务和水印任务上的性能变化情况。图 4.4 展示了微调攻击后水印 KAN 模型在原始任务和水印任务上的性能变化情况。



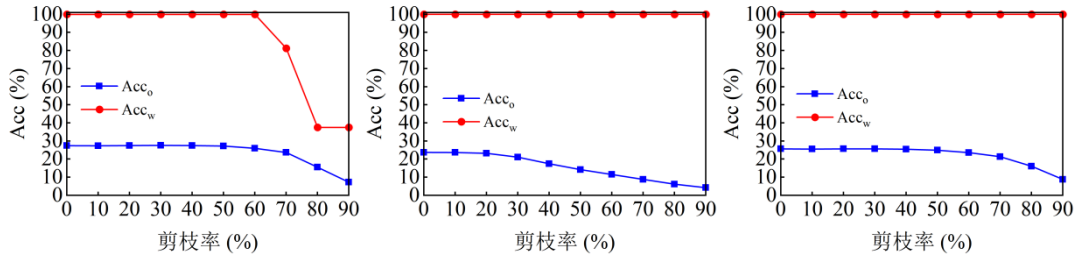
(a) Fashion-MNIST EfficientKAN (b) Fashion-MNIST FastKAN (c) Fashion-MNIST FasterKAN



(d) MNIST EfficientKAN (e) MNIST FastKAN (f) MNIST FasterKAN



(g) CIFAR-10 EfficientKAN (h) CIFAR-10 FastKAN (i) CIFAR-10 FasterKAN



(j) CIFAR-100 EfficientKAN (k) CIFAR-100 FastKAN (l) CIFAR-100 FasterKAN

图 4.3 剪枝攻击后水印 KAN 模型的性能评估

从图 4.3 中可以看出，本章所提出的水印方法在不同 KAN 模型架构下均展现出优秀的抗剪枝攻击能力。随着剪枝率的逐步提升，水印 KAN 模型在原始任务上的分类准确率呈现下降趋势，但其在水印任务上的分类准确率始终维持在较高的水平，表现出良好的稳定性。具体而言，在剪枝率低于 70%的范围内，KAN 模型在所有测试数据集上的水印任务的分类准确率几乎均能够维持在 100%，同时 KAN 模型在原始任务上的分类准确率也能够保持在较高的水平，这表明本章所提出的 KAN 模型水印嵌入方案能够有效抵抗较高强度的模型剪枝攻击。当剪

枝率提升至 70%及以上时，部分 KAN 模型的水印分类准确率开始出现下降，但此时模型在原始任务上的分类准确率也开始大幅衰减，说明攻击者若想要试图通过剪枝攻击来破坏模型水印，将会导致模型丧失基本的分类能力，从而使攻击失去实际意义。因此，尽管在高剪枝率下模型性能和水印认证效果会受到一定的影响，但本章所提出的方案仍具备较强的实用价值和安全性，在抵抗剪枝攻击方面展现出良好的鲁棒性。

(a) Fashion-MNIST EfficientKAN (b) Fashion-MNIST FastKAN (c) Fashion-MNIST FasterKAN

(d) MNIST EfficientKAN (e) MNIST FastKAN (f) MNIST FasterKAN

(g) CIFAR-10 EfficientKAN (h) CIFAR-10 FastKAN (i) CIFAR-10 FasterKAN

(j) CIFAR-100 EfficientKAN (k) CIFAR-100 FastKAN (l) CIFAR-100 FasterKAN

图 4.4 微调攻击后水印 KAN 模型的性能评估

图 4.4 展示了微调攻击后水印 KAN 模型在原始任务和水印任务上的性能变化情况。可以看出，随着微调比例的逐步提升，水印 KAN 模型在原始任务上的分类准确率没有出现明显的下降，甚至在一定程度上略有提升，而其在水印任务上的分类准确率同样始终维持在较高的水平。这一实验结果表明，攻击者难以通过模型微调攻击破坏已经嵌入的水印信息，同时也印证了本章所提出的方法在面对微调攻击时具备较强的鲁棒性。

表 4.3 展示了经剪枝-微调攻击和覆盖攻击后水印 KAN 模型的性能评估结果。结合表 4.1 可以发现，本章所提出的水印方法在两种攻击场景下均展现出优秀的鲁棒性，水印任务的分类准确性在所有测试条件下均维持在 100%，充分验证了方案的有效性和可靠性。

表 4.3 剪枝-微调攻击和覆盖攻击后水印 KAN 模型的性能评估

数据集	KAN 模型	剪枝-微调攻击		覆盖攻击	
		Acc_o (%)	Acc_w (%)	Acc_o (%)	Acc_w (%)
Fashion-MNIST	EfficientKAN	88.98	100	88.62	100
	FastKAN	88.66	100	89.64	100
	FasterKAN	89.54	100	89.13	100
MNIST	EfficientKAN	98.30	100	97.35	100
	FastKAN	97.92	100	97.30	100
	FasterKAN	98.64	100	97.79	100
CIFAR-10	EfficientKAN	55.06	100	55.50	100
	FastKAN	46.54	100	53.99	100
	FasterKAN	51.58	100	53.32	100
CIFAR-100	EfficientKAN	27.34	100	27.49	100
	FastKAN	17.26	100	23.25	100
	FasterKAN	25.68	100	25.59	100

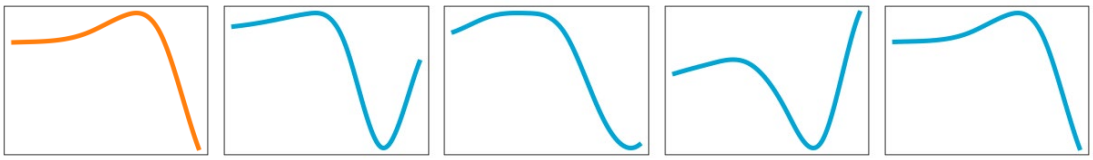
具体而言，在 Fashion-MNIST 和 MNIST 数据集上，经过剪枝-微调攻击和覆盖攻击后，KAN 模型在原始任务上的分类准确率下降幅度极小，均能够控制在 1%左右，基本保持了嵌入水印后的 KAN 模型的基准性能。同时，水印任务的分类准确性均保持在 100%。这表明对于简单数据集，这两类水印去除攻击对 KAN 模型性能的影响微乎其微。然而，在 CIFAR-10 和 CIFAR-100 数据集上，模型在原始任务上的分类准确率相较于基准水平出现了更为明显的下降。具体而言，在 CIFAR-10 数据集上，EfficientKAN 模型、FastKAN 模型和 FasterKAN 模型在剪

枝-微调攻击后的下降幅度约为 0.90%~14.53%，覆盖攻击后的下降幅度约为 0.11%~0.84%；在 CIFAR-100 数据集上，剪枝-微调攻击导致的下降幅度约为 0.22%~27.14%，覆盖攻击后的下降幅度约为 0.31%~1.86%。其中 FastKAN 模型在剪枝-微调攻击下的下降幅度最为显著，达到 14%~28%，但覆盖攻击对各模型的影响相对较小。尽管如此，水印任务的分类准确性在所有情况下仍能够保持 100%，说明即使剪枝-微调攻击和覆盖攻击对模型性能造成了一定影响，但嵌入的水印信息依然完整保留。通过对比两种攻击方式，覆盖攻击对模型性能的影响要小于剪枝-微调攻击，各数据集上模型的准确率下降幅度普遍更低，但两种攻击均未能对水印任务的分类准确性产生任何影响。因此攻击者若试图通过剪枝-微调攻击或覆盖攻击去除水印，虽可能在复杂数据集上对模型原始任务性能造成一定损害（尤其是剪枝-微调攻击），但无法破坏嵌入的水印信息，从而使攻击失去实际意义，表明所提出的方案在抵抗剪枝-微调攻击和覆盖攻击方面展现出优秀的鲁棒性与实用价值。

4.3.5 对比实验

本节对所提出的方法与现有方法[84-86]进行了对比实验，评估维度主要涵盖两个方面：一是水印嵌入前后 KAN 模型激活函数的变化差异；二是嵌入水印后的模型在四类攻击场景下的水印鲁棒性，即 80%剪枝率的剪枝攻击、80%微调比例的微调攻击、80%剪枝率与 80%微调比例依次进行的剪枝-微调攻击，以及覆盖攻击。其中，所采用的 KAN 模型为 EfficientKAN，所选数据集为 MNIST。图 4.5 展示了不同方法下 KAN 模型激活函数$\phi_{0,50,100}$的可视化对比，表 4.4 进一步给出了相应的定量分析结果。从图 4.5 可以观察到，文献[84]、文献[85]和文献[86]所生成的激活函数曲线均产生了视觉可见的明显形变；相应地，表 4.4 中这三种方法的 MAE 值分别为 0.0056、0.0025 和 0.0094，并且 PCC 值仅为 0.23、0.66 和-0.57，表明其激活函数相较于干净模型的激活函数发生了显著偏离，其中文献[86]方法的负相关性尤为突出。相比之下，本章所提出的方法生成的激活函数曲线与干净 KAN 模型的激活函数曲线几乎重合，视觉差异极小；定量结果也印证了这一观察，其 MAE 低至 0.00008，相关系数高达 0.99996，几乎与原始模型完全一

致，具有良好的形态不可感知性。这使得攻击者难以通过观察激活函数曲线的形态变化来检测水印的存在，从而有效提升了水印的隐蔽性与安全性。



(a) 干净 KAN 模型　(b) 文献[84]　(c) 文献[85]　(d) 文献[86]　(e) 本章方法

图 4.5 不同方法下 KAN 模型激活函数$\phi_{0,50,100}$的可视化对比

表 4.4 不同方法下 KAN 模型激活函数$\phi_{0,50,100}$的误差与相关性分析

方法	MAE	PCC
文献[84]	0.00559796	0.23168303
文献[85]	0.00245812	0.65613133
文献[86]	0.00937963	-0.56753433
本章方法	0.00007995	0.99996412

表 4.5 不同方法下水印嵌入前后 KAN 模型的性能

方法	嵌入水印前	嵌入水印后	
	Acc_o (%)	Acc_o (%)	Acc_w (%)
文献[84]	97.40	97.07	99.68
文献[85]	97.40	97.45	100
文献[86]	97.40	94.43	100
本章方法	97.40	97.29	100

表 4.5 展示了不同方法下水印嵌入前后 KAN 模型的性能对比，表 4.6 则为不同水印方法在各类攻击下的 KAN 模型性能对比结果。实验结果表明，四种方法在嵌入水印后，KAN 模型均能够在原始任务与水印任务上保持较高的分类准确率，说明各类方法均具备基础的水印嵌入能力。但面对不同攻击时，各方法的鲁棒性表现呈现显著差异。其中，文献[84]中的方法表现相对稳健，在剪枝、微调、剪枝-微调及覆盖攻击下，其水印任务分类准确率始终维持在 96%以上，原始任务分类准确率也能够保持在 88%以上，展现出一定的综合抗攻击能力；文献[85]中的方法在剪枝、微调与剪枝-微调攻击下，水印任务分类准确率可以保持在 90%以上，然而在覆盖攻击下，该指标骤降至 0%，暴露出其对覆盖攻击的明显

脆弱性；文献[86]中的方法鲁棒性表现最差，除微调攻击外，在其余攻击下水印任务分类准确率均大幅下滑，覆盖攻击下更是仅为 13.25%，且原始任务分类准确率也受到明显影响，可见将该方法应用于 KAN 模型时，在遭受攻击后难以保证水印的有效检测。相比之下，本章提出的方法在所有攻击场景下，水印任务分类准确率均实现 100%，同时原始任务分类准确率稳定在 95.42%~98.20%之间，不仅水印任务表现优于文献[84]中的方法，原始任务准确率也全面超越文献[84]和文献[86]中的方法，且在多数攻击场景下优于文献[85]。尤其在最具挑战性的覆盖攻击下，本章方法仍能保持 97.35%的原始任务分类准确率与 100%的水印任务分类准确率，而文献[85]和文献[86]中的方法在此攻击下水印任务性能已近乎完全失效。上述结果表明，本章方法在实现水印高保真嵌入的同时，具备极强的抗剪枝、抗微调、抗剪枝-微调及抗覆盖攻击能力，有效解决了现有方法在复杂攻击场景下水印易失效、模型性能易受损的问题，显著提升了水印技术的鲁棒性与实际应用性。

表 4.6 不同方法在各类攻击下的水印 KAN 模型性能评估

方法	剪枝攻击		微调攻击		剪枝-微调攻击		覆盖攻击	
	Acc_o(%)	Acc_w(%)	Acc_o(%)	Acc_w(%)	Acc_o(%)	Acc_w(%)	Acc_o(%)	Acc_w(%)
文献[84]	88.84	99.42	97.40	99.32	94.64	99.02	94.67	96.32
文献[85]	95.19	90.00	97.60	100	96.05	90.00	97.41	0
文献[86]	82.25	37.35	94.73	99.45	87.84	42.75	93.48	13.25
本章方法	95.42	100	98.20	100	97.60	100	97.35	100

4.4 本章小结

针对 KAN 模型水印嵌入过程会直接影响激活函数的形态而导致水印隐蔽性不足的问题，本章提出了一种基于激活扰动的 KAN 模型水印方法。该方法采用两阶段训练策略，仅在训练完成的干净 KAN 模型第一层的激活函数上施加微小扰动即可完成水印嵌入，同时引入参数对抗训练机制，通过对激活函数添加随机噪声模拟水印去除攻击场景，有效提升了水印的形态不可感知性和鲁棒性。为验

证所提出的方法的有效性，本章从保真度、形态不可感知性和鲁棒性三个维度对所提出的方法开展实验，并与现有水印方法进行对比分析。实验结果表明，所提出的方法能够在不显著影响 KAN 模型原始任务性能的前提下，稳定实现水印的嵌入与提取，具备较高的保真度与形态不可感知性，同时在多种典型水印去除攻击下仍能可靠提取水印，表现出优异的鲁棒性，在 KAN 模型安全和版权保护等应用场景中展现出良好的实际应用前景。

第五章　总结与展望

5.1 总结

随着人工智能技术的快速发展，深度神经网络已被广泛应用于计算机视觉与自然语言处理等领域，高性能模型的研发通常需要大量的标注数据、精巧的架构设计及庞大的计算资源，使其成为极具价值的数字资产，然而其在实际应用中的广泛普及也面临严峻的未授权访问风险，保护模型知识产权因此成为一项紧迫且重要的研究课题。作为一种新兴架构，KAN 模型将可学习的激活函数放置于网络“边”上而非将固定的激活函数置于“节点”处，这一创新设计显著提升了其拟合复杂函数的能力，也赋予其截然不同的结构特性。现有研究已将数字水印技术引入卷积神经网络、循环神经网络等主流架构的版权保护中，能够在原始任务与水印任务间取得较好平衡，但直接将现有方案应用于 KAN 模型时，仍然面临鲁棒性不足、隐蔽性欠缺以及嵌入成本较高等问题。针对上述挑战，本文开展了适用于 KAN 模型的鲁棒水印技术研究，取得的主要成果如下：

1）由于 KAN 模型的激活函数善于拟合连续平滑函数，其在训练过程中会优先拟合样本量大、损失下降明显的原始任务，从而导致水印任务的学习能力存在不足。同时，由于 KAN 模型的可调参数分布于激活函数上，导致 KAN 模型的参数冗余空间较传统模型更小，使得水印嵌入的难度增大。针对这一问题，本文利用频域扰动和交替训练实现黑盒 KAN 模型水印。该方案利用图像频域特征的全局关联特性，将水印嵌入频域以构建触发集，从而实现水印的均匀分布与全局影响，同时采用交替训练策略，轮流使用原始训练集与触发集对模型进行交替训练，使模型在学习原始任务的同时充分习得水印任务，并在训练过程中引入模型参数扰动损失来模拟攻击引发的参数变化，从而增强模型对剪枝、微调等攻击的鲁棒性。实验结果表明，该方案在保持 KAN 模型原始任务性能的前提下，即使面对 50%的剪枝率，水印任务分类准确率仍可达 100%；在三种不同频域触发集上遭受微调攻击后，水印任务平均性能仍然可以保持在 98.67%以上，展现出较强的鲁棒性。

2）由于 KAN 模型的参数分布于各边上的可学习激活函数，其水印嵌入过程会直接体现为激活函数形态的改变，且现有方案在训练过程中主要以水印验证准确率和原始任务精度为目标，缺乏对激活函数形变幅度施加有效约束，致使水印 KAN 模型与干净 KAN 模型在对应边上的激活函数存在显著形态差异，严重降低水印隐蔽性。针对这一问题，本文提出了一种基于激活扰动的 KAN 模型水印方案。该方案利用 KAN 模型激活函数的可解释性与结构可调特性，采用两阶段训练策略，仅在完成原始任务训练的干净 KAN 模型第一层激活函数上施加微小扰动来完成水印嵌入，避免激活函数的形态发生明显改变，同时引入参数对抗训练机制，在训练过程中对激活函数施加随机噪声以模拟攻击扰动，确保水印验证网络在模型遭受攻击后仍能稳定提取水印信息。实验结果表明，该方案在保持 KAN 模型原始任务性能的同时，能够实现水印的高保真嵌入与可靠提取，嵌入水印后的激活函数与嵌入水印前的激活函数在形态上高度相似，具备良好的形态不可感知性，且在剪枝、微调等多种典型攻击下仍能稳定提取水印，展现出优异的鲁棒性与良好的实际应用前景。

5.2 展望

本文针对 KAN 模型的知识产权保护问题，分别从提升水印鲁棒性与增强水印隐蔽性两个角度出发，提出了基于频域扰动和交替训练的 KAN 模型水印方案以及基于激活扰动的 KAN 模型水印方案，在保障 KAN 模型原始任务性能的前提下，有效提升了水印的抗攻击能力与形态不可感知性。然而，KAN 模型作为一种新兴架构，其知识产权保护仍然处于探索阶段，未来可以从以下几个方面进一步深入研究：

1）面向多样化攻击的鲁棒水印机制研究。本文主要针对剪枝、微调等常见攻击开展了水印鲁棒性优化，然而实际应用中 KAN 模型可能面临更复杂的攻击方式，如迁移学习攻击、模型窃取攻击等。未来可以进一步探索能够抵御多种攻击类型的 KAN 模型水印框架，提升水印在复杂环境下的生存能力。

2）可解释性与安全性兼顾的水印验证机制。KAN 模型因其激活函数可解释

性而具备良好的结构透明性，为水印的验证提供了新的思路。未来可以结合可解释性分析方法，设计更加透明、可审计的水印验证机制，在保障 KAN 模型安全的同时提升水印的可信度与可追溯性。

3）面向多场景的 KAN 模型水印标准化研究。随着 KAN 模型在偏微分方程求解、量子计算、时间序列预测等领域的逐步应用，其知识产权保护需求将更加多元化。未来可以围绕不同应用场景对安全性、隐蔽性、效率等方面的差异化需求，探索标准化的 KAN 模型水印技术体系，为 KAN 模型的安全部署与合规使用提供支撑。

参考文献

[1] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[2] Xu M, Yoon S, Fuentes A, et al. A comprehensive survey of image augmentation techniques for deep learning[J]. Pattern Recognition, 2023, 137: 109347.

[3] Lauriola I, Lavelli A, Aiolli F. An introduction to deep learning in natural language processing: models, techniques, and tools[J]. Neurocomputing, 2022, 470: 443-456.

[4] 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. 2015: 2722-2730.

[5] 张颖君, 陈恺, 周赪, 等. 神经网络水印技术研究进展[J]. 计算机研究与发展, 2021, 58(5): 964-976.

[6] Nikolaidis N, Pitas I. Copyright protection of images using robust digital signatures[C]//IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings. IEEE, 1996, 4: 2168-2171.

[7] Lee C H, Lee Y K. An adaptive digital image watermarking technique for copyright protection[J]. IEEE Transactions on Consumer Electronics, 1999, 45(4): 1005-1015.

[8] Zafeiriou S, Tefas A, Pitas I. Blind robust watermarking schemes for copyright protection of 3D mesh objects[J]. IEEE Transactions on Visualization and Computer Graphics, 2005, 11(5): 596-607.

[9] Podilchuk C I, Delp E J. Digital watermarking: algorithms and applications[J]. IEEE Signal Processing Magazine, 2001, 18(4): 33-46.

[10] Busch C, Funk W, Wolthusen S. Digital watermarking: from concepts to real-time video applications[J]. IEEE Computer Graphics and Applications, 1999,

19(1): 25-35.

[11] Braudaway G W. Protecting publicly-available images with an invisible image watermark[C]//Proceedings of International Conference on Image Processing. IEEE, 1997, 1: 524-527.

[12] Noorkami M, Mersereau R M. Digital video watermarking in P-frames with controlled video bit-rate increase[J]. IEEE Transactions on Information Forensics and Security, 2008, 3(3): 441-455.

[13] 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(6): 1792-1810.

[14] Li Y, Wang H, Barni M. A survey of deep neural network watermarking techniques[J]. Neurocomputing, 2021, 461: 171-193.

[15] 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.

[16] Zhang C, Bengio S, Hardt M, et al. Understanding deep learning (still) requires rethinking generalization[J]. Communications of the ACM, 2021, 64(3): 107-115.

[17] Uchida Y, Nagai Y, Sakazawa S, et al. Embedding watermarks into deep neural networks[C]//Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval. 2017: 269-277.

[18] 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 Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems. 2019: 485-497.

[19] Liu H, Weng Z, Zhu Y. Watermarking deep neural networks with greedy residuals[C]//ICML. 2021, 139: 6978-6988.

[20] Wang T, Kerschbaum F. Attacks on digital watermarks for deep neural networks[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 2622-2626.

[21] Wang T, Kerschbaum F. Riga: covert and robust white-box watermarking of deep neural networks[C]//Proceedings of the Web Conference 2021. 2021: 993-1004.

[22] Kuribayashi M, Tanaka T, Suzuki S, et al. White-box watermarking scheme for fully-connected layers in fine-tuning model[C]//Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security. 2021: 165-170.

[23] Zhao X, Yao Y, Wu H, et al. Structural watermarking to deep neural networks via network channel pruning[C]//2021 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2021: 1-6.

[24] 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.

[25] 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.

[26] Adi Y, Baum C, Cisse M, et al. Turning your weakness into a strength: watermarking deep neural networks by backdooring[C]//27th USENIX Security Symposium (USENIX Security 18). 2018: 1615-1631.

[27] Zhang J, Gu Z, Jang J, et al. Protecting intellectual property of deep neural networks with watermarking[C]//Proceedings of the 2018 on Asia Conference on Computer and Communications Security. 2018: 159-172.

[28] Guo J, Potkonjak M. Watermarking deep neural networks for embedded systems[C]//Proceedings of the International Conference on Computer-Aided Design. 2018: 1-8.

[29] Liu Y, Wu H, Zhang X. Robust and imperceptible black-box DNN watermarking based on fourier perturbation analysis and frequency sensitivity clustering[J]. IEEE Transactions on Dependable and Secure Computing, 2024, 21(6): 5766-5780.

[30] Mo M, Wang C, Bian S. A unique identification-oriented black-box

watermarking scheme for deep classification neural networks[J]. Symmetry, 2024, 16(3): 299.

[31] Le Merrer E, Perez P, Trédan G. Adversarial frontier stitching for remote neural network watermarking[J]. Neural Computing and Applications, 2020, 32(13): 9233-9244.

[32] Jia H, Choquette-Choo C A, Chandrasekaran V, et al. Entangled watermarks as a defense against model extraction[C]//30th USENIX Security Symposium (USENIX Security 21). 2021: 1937-1954.

[33] Charette L, Chu L, Chen Y, et al. Cosine model watermarking against ensemble distillation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(9): 9512-9520.

[34] Li Y, Bai Y, Jiang Y, et al. Untargeted backdoor watermark: towards harmless and stealthy dataset copyright protection[J]. Advances in Neural Information Processing Systems, 2022, 35: 13238-13250.

[35] Xu X, Li Y, Yuan C. “Identity bracelets” for deep neural networks[J]. IEEE Access, 2020, 8: 102065-102074.

[36] Zhang J, Chen D, Liao J, et al. Model watermarking for image processing networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 12805-12812.

[37] Zhao X, Wu H, Zhang X. Watermarking graph neural networks by random graphs[C]//2021 9th International Symposium on Digital Forensics and Security (ISDFS). IEEE, 2021: 1-6.

[38] 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.

[39] Fei J, Xia Z, Tondi B, et al. Supervised gan watermarking for intellectual property protection[C]//2022 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2022: 1-6.

[40] Chen H, Zhang W, Liu K, et al. Speech pattern based black-box model watermarking for automatic speech recognition[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 3059-3063.

[41] Dai L, Mao J, Fan X, et al. Deephider: a covert nlp watermarking framework based on multi-task learning[J]. arXiv preprint arXiv:2208.04676, 2022.

[42] 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.

[43] 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.

[44] Zhang J, Chen D, Liao J, et al. Exploring structure consistency for deep model watermarking[J]. arXiv preprint arXiv:2108.02360, 2021.

[45] Zhang L, Zhang X, Wu H. High-frequency artifacts-resistant image watermarking applicable to image processing models[J]. Applied Sciences, 2024, 14(4): 1494.

[46] Liu Y, Zhang L, Wu H, et al. Reducing high-frequency artifacts for generative model watermarking via wavelet transform[J]. IEEE Internet of Things Journal, 2024, 11(10): 18503-18515.

[47] Kolmogorov A N. On the representation of continuous functions of several variables by superpositions of continuous functions of a smaller number of variables[M]. American Mathematical Society, 1961.

[48] Kolmogorov A N. On the representations of continuous functions of many variables by superposition of continuous functions of one variable and addition[C]//Dokl. Akad. Nauk USSR. 1957, 114: 953-956.

[49] Braun J, Griebel M. On a constructive proof of Kolmogorov's superposition theorem[J]. Constructive Approximation, 2009, 30(3): 653-675.

[50] Haykin S. Neural networks: a comprehensive foundation[M]. Prentice Hall PTR,

1994.

[51] Cybenko G. Approximation by superpositions of a sigmoidal function[J]. Mathematics of Control, Signals and Systems, 1989, 2(4): 303-314.

[52] Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators[J]. Neural Networks, 1989, 2(5): 359-366.

[53] Liu Z, Wang Y, Vaidya S, et al. Kan: Kolmogorov-Arnold networks[J]. arXiv preprint arXiv:2404.19756, 2024.

[54] Wang Y, Sun J, Bai J, et al. Kolmogorov-Arnold-Informed neural network: a physics-informed deep learning framework for solving forward and inverse problems based on Kolmogorov-Arnold Networks[J]. Computer Methods in Applied Mechanics and Engineering, 2025, 433: 117518.

[55] Guo C, Sun L, Li S, et al. Physics-informed Kolmogorov-Arnold network with Chebyshev polynomials for fluid mechanics[J]. Physics of Fluids, 2025, 37(9).

[56] Jacob B, Howard A A, Stinis P. SPIKANs: separable physics-informed Kolmogorov-Arnold networks[J]. Machine Learning: Science and Technology, 2025, 6(3): 035060.

[57] Kundu A, Sarkar A, Sadhu A. Kanqas: Kolmogorov-Arnold network for quantum architecture search[J]. EPJ Quantum Technology, 2024, 11(1): 76.

[58] Troy W. Sparks of quantum advantage and rapid retraining in machine learning[J]. arXiv preprint arXiv:2407.16020, 2024.

[59] Genet R, Inzirillo H. A temporal Kolmogorov-Arnold transformer for time series forecasting[J]. arXiv preprint arXiv:2406.02486, 2024.

[60] Xu K, Chen L, Wang S. Kolmogorov-Arnold networks for time series: bridging predictive power and interpretability[J]. arXiv preprint arXiv:2406.02496, 2024.

[61] Vaca-Rubio C J, Blanco L, Pereira R, et al. Kolmogorov-Arnold networks (kans) for time series analysis[C]//2024 IEEE Globecom Workshops (GC Wkshps). IEEE, 2024: 1-6.

[62] Li C, Liu X, Li W, et al. U-kan makes strong backbone for medical image

segmentation and generation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2025, 39(5): 4652-4660.

[63] Cheon M. Demonstrating the efficacy of Kolmogorov-Arnold networks in vision tasks[J]. arXiv preprint arXiv:2406.14916, 2024.

[64] Ge R, Yu X, Chen Y, et al. Tc-kanrecon: high-quality and accelerated mri reconstruction via adaptive kan mechanisms and intelligent feature scaling[J]. IEEE Journal of Biomedical and Health Informatics, 2025.

[65] Liu Z, Ma P, Wang Y, et al. Kan 2.0: Kolmogorov-Arnold networks meet science[J]. arXiv preprint arXiv:2408.10205, 2024.

[66] Li Z. Kolmogorov-Arnold networks are radial basis function networks[J]. arXiv preprint arXiv:2405.06721, 2024.

[67] Delis A. FasterKAN = FastKAN + RSWAF bases functions and benchmarking with other KANs[EB/OL].(2024)

[68] Bozorgasl Z, Chen H. Wav-kan: Wavelet Kolmogorov-Arnold networks, 2024[J]. arXiv preprint arXiv:2405.12832.

[69] SS S, AR K, KP A. Chebyshev polynomial-based Kolmogorov-Arnold networks: an efficient architecture for nonlinear function approximation[J]. arXiv preprint arXiv:2405.07200, 2024.

[70] Aghaei A A. rkan: rational Kolmogorov-Arnold networks[J]. arXiv preprint arXiv:2406.14495, 2024.

[71] Samadi M E, Müller Y, Schuppert A. Smooth Kolmogorov Arnold networks enabling structural knowledge representation[J]. arXiv preprint arXiv:2405.11318, 2024.

[72] Xu J, Chen Z, Li J, et al. Fourierkan-gcf: Fourier Kolmogorov-Arnold network--an effective and efficient feature transformation for graph collaborative filtering[J]. arXiv preprint arXiv:2406.01034, 2024, 10.

[73] Blealtan A D. An efficient implementation of Kolmogorov-Arnold network[EB/OL].(2024)

[74] Ta H T. BSRBF-KAN: a combination of B-splines and radial basis functions in Kolmogorov-Arnold networks[C]//International Symposium on Information and Communication Technology. Singapore: Springer Nature Singapore, 2024: 3-15.

[75] Sundararajan D. Discrete wavelet transform: a signal processing approach[M]. John Wiley & Sons, 2016.

[76] Rao K R, Yip P. Discrete cosine transform: algorithms, advantages, applications[M]. Academic Press, 2014.

[77] Heckbert P. Fourier transforms and the Fast Fourier Transform (FFT) algorithm[J]. Computer Graphics, 1995, 2(1995): 15-463.

[78] 李屹, 魏建国, 刘贯伟. 模型剪枝算法综述[J]. 计算机与现代化, 2022, 9: 51-59.

[79] 谭景轩, 钟楠, 郭钰生, 等. 深度神经网络模型水印研究进展[J]. 上海理工大学学报, 2024, 46(3): 225-242.

[80] 马铭苑, 李虎, 王梓斌, 等. 深度神经网络模型后门植入与检测技术研究综述[J]. 计算机工程与科学, 2022, 44(11): 1959.

[81] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 2002, 86(11): 2278-2324.

[82] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images[J]. 2009.

[83] Xiao H, Rasul K, Vollgraf R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms[J]. arXiv preprint arXiv:1708.07747, 2017.

[84] Chien T Y, Shen C Y. Customized and robust deep neural network watermarking[C]//Proceedings of the 17th ACM International Conference on Web Search and Data Mining. 2024: 134-142.

[85] Kim B, Lee S, Lee S, et al. Margin-based neural network watermarking[C]//International Conference on Machine Learning. PMLR, 2023: 16696-16711.

[86] 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. 2019: 126-137.

攻读硕士学位期间取得的研究成果

[1] Zhao Y, Zhao G, Wu H. ATPP-KAN: robust black-box watermarking for Kolmogorov-Arnold Networks via alternating training and parameter perturbation[J]. Journal of Electronic Imaging, 2026, 35(1), 013004. (SCI)

致　谢

岁月流转，朝暮更迭，转眼间硕士阶段的求学之路已然行至尾声。回首这段充实而珍贵的岁月，从初入校园时的憧憬与期许，到潜心钻研、完成毕业论文的沉淀与成长，每一步都离不开师长的指引、同窗的相伴与家人的支撑。这段旅程不仅是知识积累与能力提升的过程，更是自我认知与心智成熟的修行。值此论文定稿之际，我谨以最诚挚的心意，向所有给予我关心、帮助与支持的人们，致以最深切的感谢。

首先，我要向我的导师吴汉舟老师致以最诚挚的谢意。从论文选题之初的方向指引，到研究推进中的思路点拨，再到成文过程中的细致打磨，吴老师始终以严谨的治学态度、深厚的学术素养与包容的育人情怀，为我答疑解惑、引路护航。在我陷入研究瓶颈、思路困顿之时，是您的耐心指导与适时鼓励，让我得以拨开迷雾、稳步前行。您不仅在学术上为我树立了榜样，更在为人处世、治学精神上给予了我终身受益的启迪。师恩厚重，铭记于心，在此谨向吴老师表达最深切的感激与崇高的敬意。

同时，我还要衷心感谢实验室的诸位同窗。在漫长的科研路上，正是有了彼此的陪伴与扶持，才让无数个攻坚克难的日子变得温暖而坚定。无论是实验中的思路碰撞、技术上的相互启发，还是生活里的彼此照应、困惑时的耐心解答，你们的真诚与热忱都给予了我莫大的支持与力量。与大家并肩奋斗、共同成长的时光，不仅是我研究生生涯中最为珍贵的回忆，更铸就了这段求学岁月里最明亮厚重的底色。

此外，我要将心底最深厚的感激，献给我最亲爱的家人。求学数载，远在他乡，是你们始终如一的牵挂与包容，为我筑起最安稳的港湾。在我为研究辗转难眠、倍感压力之时，是你们无言的支持与温柔的鼓励，给予我直面困难的勇气与坚持到底的底气。你们不求回报的付出、始终如一的信任，是我能够心无旁骛、潜心求学的坚实依靠。这份成长与收获，不仅镌刻着我求学路上的执着与坚守，更饱含着家人深沉的挚爱、无私的付出与长久的守护。

最后，谨向百忙之中参与本论文评审与答辩的各位专家老师，致以最诚挚的谢意。您们严谨细致的审阅、中肯专业的指点与宝贵的修改意见，为我厘清研究不足、完善论文内容提供了极大帮助，也让我在学术思考上获益良多。前路漫漫，我将带着这份启发与鞭策，继续秉持求学初心，踏实前行，不负诸位师长的教诲与厚望。

赵一飞

上海大学

2026 年 5 月 15 日