神经网络训练与推理中的机密计算技术综述：从TEE到密码学原语

王勃博 杨洪伟 郝 萌 何 慧 张伟哲
（哈尔滨工业大学网络空间安全学院 哈尔滨 150001）

摘 要 随着深度神经网络在图像识别和医疗决策等关键领域的应用日益广泛，敏感数据在训练和推理过程中频繁泄露，暴露出在特定计算环境下隐私保护机制的不足。机密计算技术通过创建数据“可用不可见”的计算环境，成为改善这些缺陷的核心手段。然而，机密计算技术的硬件局限性及其计算和通信的复杂性使得实际应用受到限制，同时，神经网络中的非线性函数也制约了基于机密计算技术的落地应用。本文系统梳理了机密计算技术在神经网络训练与推理中的各种应用路径与演进趋势，涵盖了基于SGX的多方联合训练验证、针对不同模型的性能优化与安全功能扩展，及新兴方向对机密计算应用的拓展。最后指出，尽管已有大量探索，现有技术在效率、安全性与可部署性间难以平衡，性能瓶颈、编程复杂性与标准化缺失等问题仍然制约了机密计算神经网络的规模化应用，急需在基础理论和系统工程层面实现突破。

关键词 机密计算；深度学习；同态加密；安全多方计算；神经网络；隐私计算
中图法分类号 TP309 **DOI号** 10.11897/SP.J.1016.2026.00885

Confidential Computing Techniques in Neural Network Training and Inference: From TEE to Cryptographic Foundations

WANG Bo-Bo YANG Hong-Wei HAO Meng HE Hui ZHANG Wei-Zhe
（School of Cyberspace Science, Harbin Institute of Technology, Harbin 150001）

Abstract With the extensive deployment of deep neural networks in key areas such as image recognition, natural language processing, medical decision-making and financial risk control, issues of data security and privacy protection have become increasingly prominent. During the processes of training and inference, models often need to directly access highly sensitive data, while traditional encryption and access-control mechanisms cannot effectively prevent data leakage in open or untrusted computing environments, revealing a trust gap in the practical deployment of deep learning. Confidential computing technology establishes a protected execution environment at the hardware or cryptographic level, in which data remain usable but invisible, thus providing new ideas and fundamental support for the privacy protection of neural networks. This technology enables secure model training and inference without revealing plaintext data, thereby realising trustworthy artificial intelligence across cloud, edge and multi-party collaborative settings. However, confidential computing still faces challenges such as large performance overheads, high programming complexity and limited compatibility, and its application in the field of neural

收稿日期：2025-05-01；在线发布日期：2025-11-12。本课题得到国家重点研发计划项目（No. 2023YFB4503205）、国家自然科学基金重点项目（No. U22A2036）、国家自然科学基金青年科学基金项目（No. 62202123）、国家自然科学基金面上项目（No. 62472122）、黑龙江省自然科学基金项目（No. LH2024F022）、中央高校基本科研业务费专项资金（No. HIT.NSFJG202433）资助。**王勃博**，博士研究生，主要研究领域为机密计算、高性能计算。E-mail：wangbobochn@hit.edu.cn。**杨洪伟**，博士，助理研究员，主要研究领域为隐私计算、联邦学习。**郝 萌**，博士，副教授，主要研究领域为高性能计算。**何 慧**，博士，教授，主要研究领域为信息安全、系统结构。**张伟哲**（通信作者），博士，教授，长江学者，中国计算机学会（CCF）杰出会员，主要研究领域为信息安全、系统结构。

networks requires a balance between security, efficiency and deployability. This paper systematically reviews the core technological pathways and evolutionary trends of confidential computing in neural-network training and inference, providing a comprehensive overview ranging from hardware-based security schemes built on trusted execution environments to software-based secure computation frameworks grounded in cryptographic primitives. At the hardware level, the paper focuses on the system architectures and security mechanisms of confidential-computing technologies such as Intel SGX, AMD SEV and NVIDIA GPU privacy extensions in supporting deep-learning tasks, and compares performance-optimisation strategies and system-design differences between CPU and GPU environments. From the cryptographic perspective, this paper discusses in depth the main homomorphic-encryption schemes used in neural-network inference, analysing their issues in numerical precision, latency and ciphertext expansion, as well as secure multi-party computation applied in joint training and distributed inference, examining their communication complexity and fault-tolerance mechanisms; it also introduces the latest explorations of zero-knowledge proofs in verifiable model computation and privacy-preserving inference. Furthermore, the paper reviews hybrid frameworks combining homomorphic encryption and multi-party computation, elucidating their design trade-offs and representative applications in terms of efficiency, security and scalability, and explores the latest advances in improving confidential-computing performance through hardware acceleration and system-level optimisation. Finally, the paper summarises the currently available confidential-computing framework systems for artificial neural networks, compares and analyses their advantages and limitations in terms of security models, functional capabilities, system performance and applicable scenarios, and proposes strategies and considerations for framework selection in different application scenarios including training, inference, federated learning and cross-domain collaboration. The comprehensive analysis indicates that although confidential computing has made significant progress in both theoretical research and engineering practice, existing solutions still find it difficult to achieve a full balance among efficiency, security and generality in the context of rapidly increasing model complexity, hardware constraints and growing demands for multi-party trust collaboration. Future development requires breakthroughs through the collaborative design of basic theory, cryptographic protocols and hardware architectures, and the promotion of standardized interfaces, verifiable execution and programmable security systems, so as to build a highly efficient and trustworthy ecosystem for next-generation secure intelligent computing.

Keywords confidential computing; deep learning; homomorphic encryption; secure multi-party computation; neural network; privacy computing

1 引 言

随着深度神经网络(Deep Neural Network，DNN)在图像识别、自然语言处理和医疗辅助决策等关键任务中取得广泛应用，神经网络(Neural Network，NN)训练与推理对海量高敏感数据的依赖也日益凸显[1-2]。在现实部署中，数据往往分布在不同组织之间，直接暴露原始数据以集中训练或云端推理的方式，已无法满足对数据隐私、合规性与信任边界的日益严苛要求[3-4]。尤其在医疗健康、金融科技、智能政务等涉及敏感数据的场景中，如何在保障数据机密性和完整性的同时实现有效的模型训练与推理，成为亟须解决的关键问题。传统的安全防护手段多集中于数据存储和传输阶段，然而一旦数据进入计算过程，便暴露于潜在的安全威胁之下。

本文中，机密计算(Confidential Computing，CC)是指在数据使用阶段提供隐私保护的计算机

制，包括但不限于通过硬件隔离、密码学技术（如同态加密、安全多方计算、零知识证明）或其组合方式实现对数据、模型和计算过程的端到端保密与验证。其核心在于确保数据在运行时不可见，而非仅限于依赖特定硬件的实现路径。此外，尽管部分系统研究不直接采用传统可信执行环境（Trusted Execution Environment，TEE）或密码学原语作为核心计算机制，但它们通过编译优化、异构计算、硬件加速等手段，为机密计算模型提供性能保障与工程支撑，因此同样被纳入本综述中。这些工作共同构成了从“理论安全”走向“系统可用”的完整研究路径。机密计算技术不依赖数据加密静态存储或传输路径的安全，而是在运行时主动保护模型与数据的每一环节，为神经网络的安全部署提供更强的隐私保障。

从系统架构角度来看，CCNN的典型设计可分为三个核心组件：数据输入端保护、神经网络核心计算单元保护与输出结果保护。输入端可通过加密上传、远程验证、数据预处理隔离等方式进行处理；神经网络主体结构（如卷积层、激活层、注意力机制等）则被部署在TEE环境、密文计算框架或混合信任架构中，实现隐私感知的计算执行；输出阶段则通过可信验证机制或零知识证明协议确保结果的正确性与完整性。

尽管多种CCNN方案在形式上各异，但其设计目标一致：在不破坏神经网络表达能力的前提下，最大限度地提升数据与模型的安全性。TEE系统倾向于直接执行原始神经网络结构，优势在于兼容性强、延迟低，但受限于资源瓶颈与侧信道风险；SMPC和HE等密码学方案则以更强的安全保障为核心，但常需要重构计算图、近似激活函数、重定义算子，造成计算与通信成本激增。混合方案试图在二者之间寻求平衡，通过“信任分区”或“计算分层”等策略提高可用性。此外，不同方案在对输入隐私、模型机密性、模型验证与推理可信度方面的侧重点也不尽相同，展示出机密计算在神经网络任务中的多样性演化趋势。

在CC技术的支持下，NN在多个应用场景中展现出巨大潜力。典型场景包括：

（1）隐私保护训练（Privacy-Preserving Training）：多个数据持有方联合训练一个NN模型，各自数据不可被他方窥见[5-6]。

（2）隐私保护推理（Privacy-Preserving Inference）：用户希望在本地数据不外泄的前提下，利用云端强大模型进行推理计算[7-8]。

（3）联邦学习（Federated Learning，FL）增强：在传统FL基础上，通过引入CC机制进一步增强对中间参数、梯度信息的保护[9]。

（4）跨域协作应用：如不同医疗机构间共享数据建模，以突破单一机构数据量不足的问题，同时严格遵守数据合规要求，如《通用数据保护条例》（General Data Protection Regulation，GDPR）、《健康保险可携性和责任法案》（Health Insurance Portability and Accountability Act，HIPAA）等[10]。

（5）智能合约与区块链结合场景：利用零知识证明（Zero-Knowledge Proof，ZKP）、TEE等技术，在链上或链下执行神经网络模型推理或验证，保障交易隐私[11-12]。

根据保护机制和技术实现层次，基于机密计算的神经网络（Confidential Computing-based Neural Network，CCNN）训练和推理技术大致可以划分为以下四类（如表1所示）：

表1　机密计算技术分类

技术实现层次	具体技术	简要分析
硬件级保护	Intel SGX，AMD SEV，ARM TrustZone等	利用硬件支持的TEE在CPU内创建受保护的执行空间
密码学原语	SMPC、HE、ZKP	通过纯软件协议在不可信环境中实现数据保密性和计算正确性保障
混合架构实现	TEE+SMPC，TEE+HE等的机密推理、训练、模型保护	利用TEE提升协议效率或实现可信初始化，融合各自优势，实现对敏感数据或模型进行推理/训练而不泄露内容
系统与应用构建	编程模型、框架、调度器、安全操作系统	解决部署、开发、运维过程中机密计算的实用难题

硬件级保护：以可信执行环境为代表，如Intel SGX、AMD SEV和ARM TrustZone，通过硬件隔离保障执行过程的完整性与保密性，具备低延迟、高吞吐、对现有应用较友好的特性，因而被许多神经网络推理系统选择，一些专用加速芯片（如TPU）也开始支持对隐私计算的原生加速[13]。然而，TEE技术也存在内存容量受限、容易受到侧信道攻击、硬件可信性假设较强等问题，限制了其在大规模模型训练与推理中的应用[14]；

密码学原语：包括同态加密（Homomorphic

Encryption，HE)(全同态与部分同态)、安全多方计算(Secure Multi-Party Computation，SMPC)(基于秘密分享(Secret Sharing，SS)、混淆电路(Garbled Circuits，GC)[15]等)、ZKP等方法。基于密码学的方案则无需依赖特定硬件，依托数学构造来保障计算过程中的数据隐私[16]。SMPC允许多个参与方在各自数据保持私密的前提下，协同完成计算任务，广泛用于联合训练和分布式推理。HE则允许对密文直接进行运算，使得服务器在不解密数据的情况下完成神经网络推理操作。ZKP可进一步提供计算正确性的可验证性。尽管密码学方法在安全性上更为稳固，但其计算开销巨大、通信成本高昂，导致实际应用中存在效率瓶颈；

混合架构实现：混合架构技术正在成为CC领域实现隐私保护的重要方向。通过将不同类型的隐私增强技术(Privacy-Enhancing Technologies，PETs)——如TEE、SMPC、HE、ZKP以及差分隐私(Differential Privacy，DP)等——有机结合，混合架构能够在性能、灵活性与安全性之间取得更加优越的平衡[17]。与单一技术路线相比，混合架构不仅可以根据应用场景动态优化计算资源与信任假设，还能有效应对多样化的攻击威胁与复杂的数据生命周期保护需求，但可能存在不同技术之间转换的安全性问题和资源与效率问题[18]；

系统与应用构建：在基础技术不断成熟的推动下，越来越多研究聚焦于将CC能力集成进可落地的系统框架与具体应用场景之中。该方向不仅关注底层协议的实现效率，更强调系统级的协同设计与工程优化，涉及诸如密态神经网络执行框架、跨域数据协同训练平台、隐私推理即服务(Privacy Inference as a Service，PIaaS)系统以及支持高性能推理的中间表示转换与编译优化等关键模块[19]。相关研究通常在确保安全性的前提下，综合考虑系统吞吐、通信延迟、编程易用性与可部署性，旨在推动CC从“理论可行”走向“实际可用”。此外，随着多样化应用需求的出现，如边缘部署、低带宽环境下的远程推理、多租户隐私保护等，系统与应用构建正成为CC研究中的关键实践维度。

针对在保障数据机密性和完整性的同时实现有效的模型训练与推理问题，尽管已有多种CCNN系统原型被提出，如使用Intel SGX实现的Oasis[20-21]、Slalom[22-24]，基于SMPC实现的SecureML[25]、Cryptonets[26-27]，以及结合多种技术路径的Gazelle[28-29]等，但现有研究仍面临若干重要挑战：

• 平衡计算效率与隐私保护强度的挑战：在CCNN应用中，既要保障数据、模型训练及推理过程的高度隐私性，又要实现可接受的响应时间和吞吐量。然而，在TEE技术中，虽然能通过硬件隔离快速执行原生指令，但受限于飞地(exclave)内部资源(如内存大小、可用指令集)和I/O频繁切换开销，复杂神经网络(特别是具有大量参数与激活状态的Transformer等模型)难以在Enclave内部高效运行。中间层频繁出Enclave(如访问外部内存)会打破保护边界，降低整体隐私性。在密码学技术中，如SMPC和HE，在加密状态下执行非线性激活函数(如ReLU、Sigmoid)需要复杂协议(如SS下的比较、近似多项式逼近等)，而这些操作需要大量加密运算或通信，导致训练或推理延迟成百上千倍增长[30-32]。因此，如何在保证严密隐私保护的同时，减少协议复杂度、降低通信量、提高本地计算并行度，是当前必须攻克的核心挑战。

• 模型复杂性支持不足带来的挑战：当前大部分CCNN系统，能够支持的模型通常局限于简单结构，如浅层全连接网络、低深度卷积神经网络(Convolutional Neural Network，CNN)。对更复杂的深度学习(Deep Learning，DL)模型(如BERT、ViT、Diffusion Models等)的支持非常受限[33-34]。这主要源自两方面原因：其一，复杂运算操作(如动态控制流、变长输入处理、注意力机制、归一化操作)在密文域、Enclave受限环境中实现非常困难，例如，同态加密下动态索引(如Self-Attention的Query-Key匹配)需要完全重构计算图，且开销巨大[35]；其二，DNN涉及大规模参数矩阵与中间状态的读写，内存使用远超典型Enclave可承载范围，即使分层分块处理，频繁的数据分页和验证过程也引发额外性能损耗，并暴露新的隐私风险。大规模模型训练涉及复杂的优化动态(如自适应学习率、梯度裁剪)，在机密环境下很难保持灵活性与正确性[36-37]。因此，如何对神经网络运算表达能力进行重新建模(如开发密文友好算子)、构建适配CC特性的模型结构，以及实现跨层高效内存管理机制，是当前模型复杂性支持面临的核心挑战。

• 兼顾可扩展性与部署通用性的挑战：现有CCNN系统通常针对特定任务、特定平台进行高度优化，这导致三方面问题：一，系统缺乏模块化设计，无法适配不同规模、不同应用场景下的变化需求[38]。二，通用性差，跨平台移植困难，开发者需要

针对不同硬件（如不同版本的SGX、TrustZone设备）手动调整数据布局、访问策略等；三，大规模分布式场景下（如FL、边缘协作）难以实现统一的隐私保障标准和高效协调机制[9,33,39]。因此，如何在密文计算环境下实现任务迁移、负载均衡和故障恢复，是当前基于CC的NN框架的应用性挑战。

已有研究如Mo等人在2024年的综述工作[14]，对机密计算在机器学习中的应用进行了系统性梳理，重点关注了可信硬件与基本安全机制等方面，具有重要参考价值。相比之下，本文通过广泛查阅并精读逾百篇国际前沿文献，突破现有综述仅聚焦单一技术维度的局限，构建了“理论—技术—应用”三位一体的分析体系，系统整理了机密计算与神经网络结合领域的研究演进路径。具体而言，本文在两方面进行了拓展：一是纳入了近一年出现的多项新技术与研究成果，如GPU TEE、zkPoT、混合型安全架构等，补充了文献[14]尚未涵盖的重要进展；二是覆盖范围更为全面，不仅包含底层安全机制（如TEE与密码学原语），还深入分析了系统实现、编译优化与工程落地策略，围绕神经网络训练与推理阶段的实际需求进行结构化讨论。本文所构建的研究框架，力求为理解该领域的技术演进与未来趋势提供更系统、更前沿的知识参考。本文的具体贡献有：

• 关键挑战的提炼：本文深入剖析现有研究工作在资源效率-隐私平衡、模型复杂性支持、系统可扩展性等方面的核心瓶颈，分别从硬件支持和软件协议的局限，指出了现有技术方案在突破这些瓶颈时所面临的挑战；

• 技术分类框架的构建：将基于机密计算的神经网络技术划分为硬件级保护、密码学原语、混合架构、系统与应用构建四大层次，清晰揭示不同技术路径的底层逻辑与演进关系。围绕不同技术路径下的计算模型、协议设计、系统架构与工程实现展开分析，重点探讨当前主流方案在效率、安全性与可部署性之间的权衡；

• 前沿趋势的前瞻性总结：基于对文献的深度分析，针对当前医疗、金融、政府等众多领域的隐私保护需求和基于机密计算的神经网络训练及推理能力之间不匹配的矛盾，结合当前技术手段和发展状况，提炼出了未来的研究重点和方向。

CCNN的训练与推理旨在在不可信环境中执行深度学习任务的同时，保障数据与模型的整体安全性。按照通行惯例，本文将其安全目标划分为两类：基本目标包括输入数据与模型参数的机密性、计算过程的正确性与完整性；增强目标则包括在特定场景下对计算结果的可验证性以及对模型本身的隐私保护性。本文讨论的敌手模型包括仅窃取信息的诚实但好奇者、具备篡改能力的主动攻击者，以及依赖部分可信基础设施（如TEE或加密通道）的受限敌手模型。上述安全目标与敌手假设共同构成了CCNN方案设计与技术选型的基本前提。

为统一评估不同研究方案在性能、安全性与可部署性方面的差异，本文将计算开销、通信开销与精度损失划分为“高/中/低”三个等级，判定标准如下：

• 计算开销：本文采用“相对于明文模型的性能开销倍数”作为量化指标。若计算时间/资源消耗超过明文基线模型的100倍以上，或仅能在专用硬件（如全同态加密加速器）上运行，则标记为“高”；资源消耗在10-100倍则标记为“中”；在10倍以内或能以原生SGX等通用平台高效运行者为“低”；

• 通信开销：以GB级或更高传输量、频繁远程交互者为“高”；MB级交互量为“中”；通信仅在初始化或验证阶段出现的方案为“低”；

• 精度损失：推理准确率下降$>5\%$为“高”，1%-5%为“中”，$<1\%$或无明显影响为“低”。

如图1所示，本文剩余部分组织如下：第2节分析了现有基于硬件保护的CCNN技术的原理、发展和成果；第3节阐述了基于密码学原语的CCNN的技术细节以及发展成果；第4节着墨于现有混合架构的技术和性能详情；第5节对硬件支持下的CCNN系统优化进行了阐述；第6节整理了现有流行的CCNN框架和系统，并针对不同系统和不同应用场景，作出了系统性的比较和选型分析；第7节总结了本文，并对CCNN需求和技术的未来发展方向进行了展望。

2　基于硬件的CCNN

随着对数据隐私和计算安全需求的不断提升，TEE逐渐成为CC领域中一条具有现实可部署性的重要技术路径[32,40]。早期的TEE研究主要集中于CPU层面的硬件隔离机制，以Intel SGX、AMD SEV和ARM TrustZone等技术为代表。这类方案相较于HE或SMPC具备更高的性能和更低的通信开销，因而被广泛用于保护神经网络训练与推理过程中的敏感数据。近年来，随着深度学习任务对GPU加速的依赖不断增强，GPU侧的机密计算能力（例如NVIDIA Hopper架构所引入的CC机制）也开始受

图 1 本文组织结构

到关注。虽然GPU TEE仍处于快速发展阶段，但其在体系结构和内存层级上与CPU存在显著差异，使得在GPU上实现可信执行环境面临新的技术挑战与设计权衡。本节在介绍CPU TEE的基础上，将进一步简要探讨GPU上机密计算的发展趋势。

2.1 CPU上的TEE

2.1.1 基于SGX技术的NN

随着TEE技术的发展，众多基于硬件的技术被相继提出。其中，为了应对多处理器系统中指令执行原子性不足的问题，英特尔公司提出了英特

尔®软件防护扩展（Software Guard Extensions，SGX）[41]。SGX引入了安全关键的处理器内部状态，这些状态可以在特权指令与用户模式指令之间共享[42]。通过一组专用CPU指令集，SGX允许应用程序在自身地址空间中创建被称为“Enclave”的安全容器，为其中的代码和数据页提供了强有力的完整性和保密性保障。一旦数据页被置于飞地中[43]，不可信的实体（包括操作系统及其他底层软件）无法访问这些内存区域。SGX使开发者能够将安全敏感的应用组件隔离出来，确保计算过程在不可信环境中的安全性，增强抵御漏洞攻击的能力[44]。

在此基础上，TEE技术逐步被探索应用于多方数据联合训练领域。作为早期尝试之一，文献[45]利用Intel SGX构建了受保护的训练执行环境，并提出访问模式混淆机制，抵御侧信道攻击泄露数据结构信息。具体而言，该工作要求CNN等算法具备数据无关性，即内存引用、磁盘访问和网络访问的序列均不依赖于机密数据。文献中针对支持向量机、矩阵分解、神经网络、决策树和K均值聚类，提出了数据无关的机器学习算法，提供了类似于纯密码学方案的强保密性保障：攻击者即使观察输入/输出操作序列（包括操作地址及加密内容），也无法区分两个大小一致、输出结果大小一致的数据集。该工作首次验证了TEE在多方协作学习中的可行性，并结合远程证明确保了执行环境的可信性。然而，其支持的模型类型较为有限，在神经网络上尚不具备实际可行性，且存在较高的性能开销。

沿着这一方向，文献[46]以[45]为基础，进一步探索了在云端提供受保护DNN训练服务的可能性。该工作结合SGX与TensorFlow，提出了Enclave封装机制，使得模型所有者能够将训练代码部署在云端，同时实现输入数据与模型参数的端到端保护。系统支持使用封装后的TensorFlow进行神经网络训练，并通过SGX实现多租户数据保护。但受限于SGX EPC容量（128 MB），在训练大规模模型时遇到了严重的页交换瓶颈，同时访问模式保护尚未完善。

由于NN推理任务的计算量相对于训练任务来说小得多，因此，随着TEE技术对NN支持能力的发展，众多研究工作首先从推理阶段的隐私与性能问题入手。文献[47]基于SGX技术提出了一个针对模型评估时的公平性问题的公平性审计框架，旨在解决数据隐私、模型保密性和可信性等潜在的安全问题。通过安全Enclave远程证明原语结合公开审查和先进的基于软件的安全技术构建信任链，使公平机器学习模型能够被安全认证，并允许客户端验证已认证的模型。该研究是SGX技术在该领域的一个重要发展，为后续基于对NN训练和推理的研究提供了借鉴。进一步地，文献[22]提出在推理过程中，将DNN中计算密集的线性层（如矩阵乘）外包给不可信图形处理单元（Graphics Processing Unit，GPU），在Enclave内部仅保留非线性层，并引入随机掩码机制验证外包计算的正确性。文献[48]则针对推荐系统，推理工作，提出了一种结合近内存处理的机密计算系统设计方案，但该方案中的内存开销和推理延迟降低了推荐系统的应用可行性。围绕推理阶段的进一步优化，文献[49]揭示了DNN最佳分区配置存在数据集与模型强相关的困难。基于此，提出了TEESlice方法：在DNN推理期间对抗MS（Model Stealing）和MIA（Membership Inference Attack）。不同于现有方法，TEESlice采用先分区后训练策略，准确区分隐私相关权重与公共权重，提供了与将整个DNN模型放入TEE内相同水平的安全保护（即“上限”安全保证），同时在开销上比TSDP方案减少超过10倍，无精度损失。文献中亦提及将提供代码和相关文档，以促进实用化进展。

随着DNN加速需求的增加，针对轻量化神经网络的TEE保护亦被提出。文献[50]注意到量化神经网络（Quantized Neural Network，QNN）虽然对计算资源需求较低，但由于量化时产生的误差，导致模型准确率下降明显。因此，提出了TEE保护的QNN分区方案（TSQP），首次在QNNs上实现了模型保密性、推理完整性及模型实用性的TEE保护安全推理框架。

除常规DNN模型外，针对结构更复杂的图神经网络（Graph Neural Network，GNN）应用场景，文献[51]提出了结合SGX与局部差分隐私（Local Differential Privacy）的方法，构建隐私保护的联邦GNN训练框架。该框架允许在本地用户数据基础上进行GNN模型的分布式训练，有效克服数据量不足的问题，并实现安全的参数共享与恶意服务器防护，确保推荐系统中用户隐私不被泄露。

另一方面，在CNN推理服务领域，文献[52]针对模型即服务（Model-as-a-Service）模式下存在的数据和模型泄漏风险，以及现有隐私保护方法在安全性与性能上的不足，提出了结合TEE受限内存特性、兼容SIMD技术的高效CNN安全推理协议，为CNN计算提供了新的相对安全且高效的执行方式。

在系统层面支持方面，为了降低TEE应用部署的复杂性，GitHub上的Gramine项目应运而生，其前身为文献[53]提出的Graphene-SGX。该项目基于库操作系统（LibOS）设计，目标是使未经修改的应用程序可以直接在SGX环境中运行。尽管存在性能与TCB规模方面的挑战，Graphene-SGX以其良好的兼容性和实用性成为SGX研究的重要里程碑，为可信计算技术的普及奠定了基础。未来的研究可进一步在优化性能、减小TCB、扩展应用场景等方面持续推进。

除了x86体系外，开源硬件领域也在积极探索TEE技术[54]。基于文献[55]，文献[56]提出了基于RISC-V指令集的Keystone框架。Keystone支持多种隔离机制（如内存保护、权限管理），并允许用户自定义安全监控器（Security Monitor）、Enclave运行时及信任根。尽管在Enclave切换效率与内存加密方面尚有优化空间，且整体安全模型未经历大规模实际攻击检验，但Keystone在可定制性与开放性方面为学术界和工业界提供了新的研究平台。基于Keystone的进一步演进英特尔最新推出的TDX（Trust Domain Extensions）技术，安全操作系统内核Gramine TDX[57]和针对大语言模型（Large Language Model，LLM）推理的EncChain方案[58]已被提出，但受限于硬件兼容性限制，以及内存加解密和硬件隔离机制带来的性能开销，Gramine TDX和EncChain仍需进一步开展优化和平衡研究。

Slalom框架[22]是最早结合使用GPU和TEE的方案之一。该框架将深度神经网络中所有线性层的计算安全地委托给不可信但高性能的协处理器（如GPU），而将非线性层的计算保留在可信执行环境（如Intel SGX或Sanctum）中执行，从而支持隐私推理任务。然而，Slalom仅适用于推理阶段，且其应用范围受限于此架构设计。在此基础上，Goten框架[23]进一步扩展了Slalom的方法，从广义上讲，两者均可视为早期基于GPU TEE的应用探索。需要指出的是，其采用的低比特宽度策略可能限制其在高精度任务中的适用性。此外，也有研究尝试通过软件方式提升GPU的可信度，例如Sage框架[59]提出了面向GPU的软件级认证机制。然而，此类方案依赖于GPU软件栈的完整性保障，且通常带来较高的性能开销。

综合而言，TEE结合DL的研究范式，正沿着“从可行性验证，到应用增强，再到性能突破，最终扩展安全功能”的技术演进方向不断发展。后续如S-NN等工作，亦更加注重系统层级的适配与优化，探索在受限Enclave中深度模型的可部署性极限。

2.1.2 基于SEV技术的NN

AMD的安全加密虚拟化（Secure Encrypted Virtualization，SEV）技术，作为一种基于虚拟化层实现内存加密的TEE方案，提供了一种无需修改操作系统或应用程序即可实现虚拟机内数据隔离保护的方法[60-61]。与Intel SGX相比，SEV拥有更广阔的内存保护范围与更为友好的编程接口，使其近年来成为多个研究探索神经网络私密推理与训练的核心平台之一[62-63]。

在初期，研究主要集中于基于SEV平台实现安全推理。文献[64]提出了将PyTorch/ONNX推理运行时隔离在SEV保护的虚拟机中，同时将用户模型封装在Docker容器中部署。通过提供完整的Attestation（远程验证证明机制）流程，该方案能够保证模型在部署与运行过程中未被篡改，从而在SEV平台上实现了“开箱即用”的安全AI推理能力，为后续可信推理平台的工程化打下了基础。

随后，学者的研究焦点进一步扩展到神经网络训练任务上。文献[65]首次验证了SEV在DL训练任务中的可行性，提出了SecureTF框架。这项工作首次将AMD SEV深度集成至TensorFlow，构建了一个端到端安全的神经网络训练与推理平台。具体而言，SecureTF通过在受SEV保护的虚拟机中运行完整的TensorFlow环境，以保障模型权重与训练数据在云端不会被宿主机窃取。同时，框架引入Attestation以确保训练环境的完整性。该方案对原生TensorFlow的改动小，兼容性良好，且支持完整训练流程，在一定程度上突破了SGX中受限于Enclave大小的性能瓶颈。然而，SecureTF仍存在访问模式保护缺失的问题，对物理攻击（如冷启动攻击，如通过物理手段读取内存残留数据的攻击）存在潜在风险。此外，SEV（而非更先进的SEV-ES和SEV-SNP）版本无法保护寄存器及管理结构，这也限制了其安全性。

在神经网络训练基础上，研究进一步探索了更复杂的分布式场景。文献[66]提出了基于SEV的FL系统框架DeTA。该框架采用去中心化的聚合策略，并在设计上引入了深度防御机制。在DeTA中，各参与方将本地模型更新划分并打乱成多个随机分区，分别分配至受CC环境保护的多个聚合器。为了支持动态、多聚合器的FL生态，DeTA还实现了一个两阶段认证协议，使新参与方能够验证所有

受保护聚合器的安全性，并建立安全通道传输模型更新，从而在更复杂的协作环境中实现了隐私保护与安全保证。

在基于AMD GPU的场景中，近年来也涌现出一些结合TEE的具有代表性的研究探索。例如，Jang等人提出的异构可信平台方案[67]，通过修改CPU与GPU之间的I/O互联机制，并将GPU驱动重构至在CPU端的可信执行环境中运行，从而实现了对GPU的安全调度与访问控制。该方法在系统架构层面具有一定的通用性，但其整体安全性仍依赖于CPU侧TEE的完整性保障。此外，Honeycomb框架[68]提出在GPU程序加载阶段引入静态分析，以实现对GPU应用的安全性验证。尽管该方法在静态场景下具有较高的分析效率，但其对静态分析工具的完整性和准确性要求较高，对于包含复杂控制流或动态生成代码的程序仍面临验证能力不足的挑战。

尽管SEV技术在推理与训练场景中表现出较大潜力，但其底层加密机制亦暴露出一定局限。文献[69]系统性地指出了AMD SEV内存加密存在的根本性问题：SEV采用无状态且未经认证的加密模式，且对加密后内存密文的读取访问几乎不受限制。这些问题为提升AMD SEV的安全性提供了重要的研究方向和改进思路。

2.1.3　基于TrustZone技术的NN

ARM TrustZone作为一种硬件支持的TEE技术，通过将系统划分为安全世界与非安全世界，提供了在不可信环境中隔离和保护敏感计算任务的能力[14,70-71]。随着对安全计算需求的不断增长，TrustZone平台逐渐成为保护敏感应用的重要基础设施之一。

早期工作主要关注于在TrustZone中构建通用的安全执行环境。文献[72]提出了TrustShadow框架，在TrustZone的安全环境中搭建了一个轻量级运行时系统，以隔离并保护应用程序的执行安全。通过拦截和验证系统调用，TrustShadow有效保障了应用与非安全世界之间的交互完整性。这一工作为在TrustZone上运行未经修改的应用程序提供了可行的安全执行方案，虽然其主要面向通用应用场景，尚未针对DL任务进行专门优化。

在此基础上，研究者开始探索TrustZone在DL安全领域的应用。受到TrustShadow启发，文献[73]首次提出了在不受信任设备上利用TrustZone保护DL模型的概念。该方案将模型执行的关键部分迁移到TrustZone的安全世界中，从而防止模型被窃取或遭篡改，为敏感机器学习推理提供了新的保护途径。具体而言，该研究对模型进行划分，并按顺序执行推理，以使模型能够在TEE内完整运行，从而实现对模型窃取的全面防护。这种方法的缺点在于，TEE内的计算资源有限，因此整体推理延迟会增加。作为一种改进，DarkneTZ[74]采用了对模型层进行划分的策略，仅在TEE内执行部分敏感层，其余层则在富执行环境（Rich Execution Environment，REE）中运行。然而，某些卷积层以明文形式暴露给攻击者，这削弱了对模型机密性的保护，因为其降低了重新训练过程的难度和工作量[75]。

随着应用场景逐渐扩展到资源受限设备，进一步的研究着眼于提高TrustZone上推理执行的效率与适用性。文献[76]针对消费级物联网设备，提出了一种内存高效且安全的DNN推理方法。具体而言，该方法设计了内存优先级管理机制，以缓解内存泄漏风险及重叠冲突问题。同时，研究引入了两个轻量级库——小型DL库S-Tinylib与Tinylibm，以支持在TEE内部实现高效推理。尽管该方法主要适用于轻量级模型，可能难以直接应用于大型DL任务，但其有效验证了在资源受限物联网设备上，借助TrustZone进行安全、高效推理的可行性。

在面向Arm GPU的场景中，也有若干结合TEE的研究方案被提出。StrongBox技术[77]提供了一个隔离的执行环境，并支持Arm GPU与CPU共享统一内存地址空间，从而提高数据交换效率并简化可信计算架构。然而，尽管StrongBox为敏感计算提供了隔离保障，其在分布式计算场景下的节点通信安全问题尚未得到充分解决，同时该方案的通用性仍有待进一步提升，以支持更广泛的应用场景，如协作式卷积神经网络（CCNN）。此外，ACAI框架[78]采用系统性方法将可信计算的安全不变量扩展至设备端访问路径，针对多个关键漏洞提出了统一的解决方案。尽管该架构在安全性方面取得重要进展，但其对非线性计算的支持能力、对特定硬件设备的依赖性，以及在高负载或高并发环境下的可扩展性仍缺乏深入评估。

2.2　GPU上的TEE

随着人工智能模型规模的持续扩大，机密计算场景中的GPU已获得了越来越多的应用。与CPU相比，GPU在体系结构与执行模型上具有显著差异，尤其是在高带宽显存（HBM）访问、并行任务调度及CPU-GPU跨设备数据传输方面。传统基于

EPC/页表机制的CPU TEE难以直接适配这一高并行、高带宽的计算模式，GPU机密计算的核心问题则在于：如何在大规模并行和高速内存环境下重构可信边界，实现安全与性能的平衡。针对这一挑战，相关研究逐步形成了两条互补的技术路径——硬件层面的可信执行支持与软件层面的系统级优化。前者着重于建立独立的信任根与内存加密机制，后者则聚焦于安全与性能的协同设计，二者共同推动了GPU机密计算从实验性探索走向体系化演进。

在硬件层面，研究的重点在于构建可独立建立信任根的GPU TEE体系。HETEE[79]提出异构可信执行环境的总体框架，通过在安全控制器上集中管理GPU与其他加速器，实现任务的动态迁移与安全调度，在不修改芯片结构的前提下完成跨设备可信编排。StrongBox[80]在Arm平台构建GPU侧轻量级TEE，依托TrustZone的安全上下文与细粒度内存保护，提供可直接部署的隔离执行环境。Guardain[81]将“设备内信任根”思路扩展至专用AI加速器，在芯片内部实现任务与模型级认证、访问控制与加解密操作，覆盖模型与数据全生命周期。NVIDIA Hopper架构的推出进一步标志着GPU机密计算进入原生支持阶段[82]。该架构在设备端引入显存加密、远程证明与安全引导机制，使GPU能够独立建立信任根，并将安全边界由“CPU主导”转为“GPU自主”。与以EPC/页表为核心的CPU TEE不同，Hopper更关注高带宽显存与并行通信路径下的安全一致性：既要保证大规模并行访存中的加密透明性与数据完整性，又需在多流并发与跨设备通道中维持端到端的度量与证明。这些研究共同构成了GPU机密计算在硬件维度上的基础体系——从异构协同、设备级可信根到原生加密机制的逐步完善。

在软件层面，研究重点逐渐转向性能与安全的系统级协同。针对单GPU场景，Honeycomb[68]通过静态分析与形式化验证约束GPU内核执行，以软件验证替代部分硬件隔离，在性能、可移植性与通用性之间取得平衡。PipeLLM[83]则从系统调度角度出发，提出推测式流水化加密与异步解密机制，将加解密与计算、传输过程并行重叠，使安全操作延迟被流水线调度所掩蔽，从而在不削弱威胁模型的前提下降低性能开销。在多GPU协同与异构并行场景下，研究者主要关注如何在安全约束下提升数据迁移与通信效率。文献[84]首先提出通过动态与批处理的元数据管理机制，在多GPU环境中实现高效的内存保护与访问控制，从而显著降低跨设备验证与同步的代价。在此基础上，文献[85]进一步优化通信路径，将控制权下沉至GPU，使设备端能够直接发起分区化MPI通信，减少CPU介入导致的切换与同步延迟，从而提升系统级并行性。与此同时，基于Grace－Hopper平台的研究[86]从统一内存模型角度验证了这一类优化的性能与安全权衡，揭示任务划分与数据分布策略会直接影响加密路径与延迟传播，为异构协同下的性能—安全折中提供了量化依据。这一系列研究形成了从内存保护机制到通信架构再到系统验证的连续探索路径，共同推动了GPU机密计算向高效、安全、可扩展的方向演进。

表2汇集了代表性的硬件与系统方案及其关键属性，便于与神经网络训练与推理场景进行对照。可以看出，这些研究虽然在安全体系结构方面取得了显著进展，但仍存在若干共性挑战。无论是传统CPU方案（如SGX、SEV）还是新兴GPU方案，在隔离粒度、系统调用支持、异构调度能力以及软件栈可信性方面仍存在局限。此外，这些方案普遍依赖底层平台特性，远程认证过程的额外开销也进一步限制了它们在协作式神经网络（CCNN）等应用中的可扩展性。在此背景下，越来越多研究开始探索与硬件TEE互补的方向——即基于密码学原语的协作式神经网络（CCNN）技术。该类方法具备更强的通用性与平台独立性，可与硬件TEE形成优势互补：前者提供跨平台一致的安全语义，后者提供设备端的高效可信执行，两者结合有望在高安全与高性能兼容的场景下支持大模型训练与推理。综上，GPU上的机密计算已逐步形成从硬件可信根到系统级性能优化的完整研究体系，其在内存加密、数据迁移与多设备安全协同等方面的差异化设计，为克服传统CPU TEE在大规模并行计算中的瓶颈提供了重要思路。

3 基于密码学原语的CCNN

基于密码学原语的CC技术正广泛应用于神经网络的训练与推理中。SMPC支持多方在不泄露数据的前提下协同计算，HE允许在密文上直接执行模型运算，ZKP则保障计算结果的可验证性[87-88]。这些技术共同提升了模型训练与推理的数据隐私保护能力，促进了安全可信的人工智能（Artificial

表2　基于硬件的优秀CCNN研究一览

文献	基础技术					应用		计算开销	通信开销	精度损失	安全目标	敌手模型
	TEE	SMPC	HE	ZKP	DP	NN	FL					
[41][72]	√							低	中	低	执行完整性	被动敌手
[45]	√					√		低	中	低	执行完整性	被动敌手
[46][50][73]	√					√		高	中	中	数据机密性	受限信任模型
[22][47][74]	√					√		高	高	低	数据机密性	被动敌手
[51][52]	√				√	√		中	高	高	模型私密性	主动敌手
[53]	√					√	√	中	中	高	数据机密性	受限信任模型
[55][57]	√					√	√	高	高	低	数据机密性	受限信任模型
[64]	√					√	√	高	高	中	数据机密性	受限信任模型
[65]	√					√		高	高	低	计算正确性	被动敌手
[66]	√						√	高	高	中	数据机密性	受限信任模型
[68][83]	√					√	√	高	高	低	计算正确性	被动敌手
[76]	√					√		高	低	高	数据机密性	被动敌手
[79][80][82]	√					√	√	高	高	低	数据机密性	主动敌手
[81]	√					√	√	高	中	低	数据机密性	主动敌手
[84][85][86]	√					√	√	高	中	中	计算正确性	被动敌手

注：本表中“高/中/低”分别依据如下标准评估：计算开销指相对明文模型计算时间＞100倍为“高”，10-100倍为“中”，＜10倍为“低”；通信开销以数据量与交互频率判断；精度损失按准确率下降幅度划分。详情参考第1节。

Intelligence，AI）系统构建。HE、SMPC和ZKP是当前构建CCNN的三大主要密码学技术，各自具备不同的安全模型与系统特性。

HE支持在加密数据上直接进行计算，适合单方私密推理与模型保护等非交互场景，但面临密文膨胀、自举开销高等计算瓶颈。SMPC通过数据拆分与交互协议实现多方联合计算，适用于跨域协作训练与联合建模等高交互场景，具备较好的效率扩展性，但对通信同步与带宽条件较为敏感。ZKP则强调在不泄露数据的前提下验证计算正确性，适合合规审计、链上推理等需要可验证性的任务，但生成与验证成本仍较高，尚难广泛部署于高频神经网络任务中。

总体来看，HE强于单方私密性保护，SMPC强于多方联合计算协作，而ZKP强于结果可验证性与公开透明性。三者适用于不同的信任模型与系统需求，近年来也出现了多种混合框架，试图融合其优势以构建更高效、安全、可用的CCNN系统。

3.1　基于SMPC的NN

SPDZ协议作为一种多方安全计算（SMPC）方案，专注于为大规模计算任务提供高效的密码学保障[89]。其核心思想是采用混合密码学方法，通过结合SS与HE，实现了对大规模运算的支持。随后，SPDZ2对协议进行了进一步优化，以提升整体性能。

2012年，文献[90]首次提出了能够抵御主动攻击者、适用于n个参与者中最多$n-1$个参与者被破坏情形的通用SMPC协议SPDZ。这一突破性成果，打破了此前在不诚实多数场景下无条件安全协议难以实现的局限，同时较大程度地降低了基于计算假设的安全协议中的公钥开销问题。尽管SPDZ在预处理阶段的计算复杂度已相较于早期方案有所降低，但仍然依赖公钥机制，其计算量随参与方数量n呈平方增长，给大规模参与场景带来了一定负担。

为了进一步提升效率，文献[91]提出了MASCOT协议，引入了不经意传输（Oblivious Transfer，OT）[92-94]这一基础密码学原语，与算术运算流程深度结合，从而优化了恶意环境下的算术安全计算。MASCOT在SPDZ的基础上，专门通过设计高效的加法器和乘法器，显著降低了计算开销，从而提升了在神经网络训练场景下使用SMPC的可行性[95-97]。同时，通过减少冗余通信和精简协议设计，MASCOT有效降低了通信开销，同时具备抵御多种恶意攻击的能力。不过，由于该方案高度依赖OT，在计算资源紧张或网络条件不佳的环境中，整体性能仍可能受到一定影响。

在SPDZ与MASCOT奠定基础后，文献[98]基于这两项工作，提出了一个实际可用、抵御主动攻击的分布式密钥生成SCALE-MAMBA框架，并在SCALE-MAMBA框架内实现了分布式BGV密钥

生成。这为后续基于SPDZ体系进行NN训练与推理提供了坚实的密钥基础。基于这一成果，文献[99]进一步提出了MD-SONIC框架，通过针对矩阵乘法、ReLU与Maxpool等算子设计快速安全模块，实现了在不诚实多数及恶意敌手存在场景下，既具备高效在线性能又保证恶意安全的NN推理，显著缓解了通信、运行时开销高企及客户端负担重的问题，尤其适合资源受限设备。虽然该框架取得了重大进展，但在实用系统部署方面，仍然存在很大的资源开销。

随后，文献[100]将其进一步扩展，提出了MP-SPDZ。该系统集成了超过30种多方计算协议变体，涵盖了从诚实多数到不诚实多数、从半诚实到恶意模型的全安全模[90,101-103]，并同时支持二进制与算术电路的高效计算。MP-SPDZ通过提供统一的Python编程接口，整合了SS、HE、OT及GC等多种底层原语，极大推动了NN训练与推理应用在SMPC框架下的实现。但这些方案均存在实际通信负载高、计算开销难以在实际部署中被接受的问题。而在神经网络训练阶段的隐私保护研究中，文献[25]提出了两方计算的SecureML框架，支持秘密分享下的安全算术运算，并设计了sigmoid与softmax函数的替代实现方式。通过C++实现，SecureML能够扩展至数百万样本与千维特征的数据集，体现了良好的可扩展性，但仍需进一步降低通信负载和计算开销。

在通信效率优化方面，文献[104]提出了Cenia模型。Cenia利用算术SS，开发了低交互的安全比较协议，有效支持安全激活函数（如ReLU）和池化层（如Maxpool）计算，无需引入GC或OT，大幅减少了通信开销。此外，该工作还设计了安全指数运算与除法协议，用于实现Sigmoid等归一化层的安全计算。文献[105]基于SS技术实现了一个隐私保护系统SecureNLP，其中，重点针对基于循环神经网络（Recurrent Neural Network，RNN）的带注意力机制的序列到序列模型（用于神经机器翻译）场景。具体而言，针对sigmoid和tanh等非线性函数，使用SMPC设计了两种高效的分布式协议，以降低计算复杂度和通信负载。文献[106]提出了CoPriv框架，CoPriv框架通过基于Winograd变换[107]的新型两方计算（Two-Party Computation，2PC）卷积协议以及面向DNN特性的方法，减少了推理通信量，体现了协议与架构协同设计的趋势。文献[108]基于SecureNLP设计了SecureGPT框架（将在第4节介绍），然而，这些工作在实际应用方面仍然需要较多的计算资源和通信量。

除通信和计算负载的优化外，很多研究也聚焦于SMPC下的CCNN训练和推理工作。在两方安全计算领域，已经涌现出多项重要成果。文献[109]提出的C2PI框架，通过将神经网络前几层的计算划分为仅需使用SMPC协议处理，大大减少了总体SMPC开销，提升了推理效率，尽管以略微放宽的数据隐私保证为代价。文献[110]提出了ABNN2框架，通过结合高效矩阵乘法协议、ReLU优化协议以及任意位宽量化NN，实现了更加高效的两方NN推理。文献[111]则设计了QUOTIENT，支持离散化DNN训练，并引入了多种优化组件以提升WAN环境下训练速度。

在三方安全计算方面，文献[112]提出了端到端隐私保护训练与推理框架FALCON，基于复制SS与协议优化，有效提升了大型机器学习模型处理的效率，但框架的计算时延仍然过高。随后，文献[113]基于FALCON设计了SEPPDL框架，在GPU上实现了高效的密文比较和矩阵乘法，为大规模CCNN推理提供了重要支持，不过，该框架采用明文模型＋密文数据的方式进行计算，一定程度上面临合谋攻击的安全隐患。

为了实现SMPC与主流机器学习生态的深度融合，文献[95]提出了CrypTen框架。CrypTen基于SPDZ协议，支持GPU上的三方半诚实安全计算，结合算术与二进制SS，优化了线性与非线性计算，并通过加密张量（CrypTensor）扩展了PyTorch生态，极大地降低了实际部署成本。但该框架中采用浮点数据类型进行计算，增加了数据在浮点和整型之间的转换开销，占用了更多的存储和计算资源。文献[114]考虑到SMPC允许多方在不泄露敏感数据的前提下联合计算函数，为隐私保护机器学习（Privacy-Preserving Machine Learning，PPML）提供了可行的解决方案，但对于缺乏密码学背景的用户而言，使用SMPC技术开发高效的PPML程序是一项巨大挑战。于是，该文献提出了一个高性能且用户友好的PPML框架SecretFlow-SPU，该框架可兼容现有DL程序。该方案具备了一定的通用性，虽然其在满足通用性时舍弃了一部分模型计算效率，也增加了通信开销，但该框架是一个致力于现有明文模型向CCNN模型转换的优秀尝试。

最后，针对不同安全模型与计算任务需求的灵活性挑战，文献[115]提出了混合SMPC协议框架

ABY。该框架整合了算术SS、布尔SS与姚氏GC，为两方安全计算提供了高效方案。随后，ABY2.0[116]与ABY³[117]分别扩展至三方计算环境，设计了高效的密文转换机制，进一步丰富了隐私保护机器学习框架的应用，适用于线性回归、逻辑回归与神经网络等多种模型。

此外，二进制神经网络的安全也受到了不少关注。针对二进制神经网络（权重和激活值均为二进制的神经网络）的CC研究最初采用了基于通用2PC的技术[118-119]。虽然这些方案能够应用于二进制神经网络的隐私推理，但由于未针对二进制特性进行专门优化，存在较大的性能瓶颈。为提升效率，文献[120-121]基于GC和OT分别实现了二进制NN中的非线性与线性操作，但计算与通信开销仍然较大。进一步的，文献[122]提出了专门面向二进制神经网络的高效安全两方推理框架SecBNN，针对每一层的特性定制密码协议，充分利用了二进制特性，优化了框架性能。然而，在大规模实际部署方面，仍有待更多研究推动。

现有研究普遍关注基于SMPC的CCNN在计算与通信开销上的挑战。文献[113]从理论层面指出，非线性运算显著增加了计算与通信轮次；文献[123]则通过实验数据对该现象进行了验证。

3.2　基于HE的NN

HE作为一种允许在加密数据上直接执行运算的密码学技术，不仅能够确保数据在处理过程中的安全性，还实现了对敏感数据的有效利用，特别适用于医疗、金融等领域对隐私保护要求极高的应用场景[124-125]。随着HE技术的不断演进，其在NN推理与训练中的应用逐步拓展与深化。

3.2.1　BGV/BFV和CKKS

HE的发展最初源自对部分同态加密（Partially Homomorphic Encryption）方案的研究。1999年，文献[126]提出了Paillier加密方案，支持加法同态，为加密数据上的简单运算提供了可行性，广泛应用于电子投票等领域。然而，Paillier方案仅支持加法操作，无法实现复杂计算。为突破此限制，2009年，文献[127]提出了第一个全同态加密（Fully Homomorphic Encryption，FHE）方案，通过引入自举（Bootstrapping）技术，理论上支持在加密数据上进行任意数量的加法与乘法操作，开启了HE在通用计算领域应用的可能性。尽管如此，第一代FHE方案效率极低，密文膨胀严重，计算开销巨大，限制了其实用性。

为改善早期FHE的性能瓶颈，2012年，文献[128-129]提出了BGV方案，首次实现了无需频繁自举操作的分层全同态加密（Leveled FHE），通过模切换（Modulus Switching）与密钥交换（Key Switching）技术有效控制噪声增长。BGV基于容错学习（Learning With Errors，LWE）问题的困难性，能够在限定深度下执行同态运算，大幅提升了效率。同年，文献[130]在BGV基础上提出了BFV方案，进一步简化了参数选择与噪声管理，增强了易用性与工程实现的可行性。

在支持精确整数运算的BGV、BFV之后，文献[131]提出了CKKS方案，开创性地支持加密浮点数的近似加法与乘法运算，引入了重缩放（Rescaling）操作与复规范嵌入（Complex Embedding）批处理技术。CKKS极大扩展了HE在机器学习、DNN等领域的应用潜力。随后，诸多研究围绕CKKS展开，如文献[132]提出的PPDNN-CRP框架，结合DNN与CKKS实现了信用风险预测中的全流程隐私保护；文献[133]提出了RNS-CKKS下无需手动规模管理的性能感知静态分析方法；文献[134]提出了NeuJeans方案，基于CinS编码实现了高效卷积计算；文献[135]提出了ReSBM编译器，优化了规模管理与自举放置。此外，文献[136-139]基于CKKS进一步开发了ANT-ACE隐私推理编译器、SIMD隐私计算解决方案、安全高效的Softmax拟合函数以及AutoFHE推理框架，整体推动了CCNN推理在性能与实用性上的快速发展。

在HE理论发展逐渐成熟的同时，HE在神经网络中的实际应用也开始受到关注。文献[26]提出了CryptoNets方法，将已训练的神经网络模型转换为可在加密数据上推理的形式，首次展示了HE结合机器学习的可行性。随后，微软研究院基于此思路开发了SEAL（Simple Encrypted Arithmetic Library）库，为HE在应用层面的推广奠定了重要基础。

3.2.2　HE-Transformer与CHET

基于SEAL库，文献[119，140]进一步探索了高效CCNN计算方案。2018年，英特尔发布了开源HE-Transformer项目[1，141-142]，基于nGraph编译器，集成了SEAL和ABY库，实现了对非多项式激活函数的支持，同时支持与TensorFlow集成推断[143-145]。进一步地，文献[146]提出了专为HE优化的Transformer架构，并开发了算子多项式转换方法，推动了HE在语言模型（Language Model，LM）

和视觉Transformer(Vision Transformer,ViT)领域应用的探索研究。

为了降低HE程序开发的复杂性,文献[147]开发了特定领域编译器CHET,实现了从高层描述到FHE程序的自动化转换与优化。基于CHET,文献[148]提出了RLNet,以期减少高延迟ReLU操作,提高模型性能;文献[149]提出了CrossNet,针对延迟敏感应用优化了隐私推理流程;文献[150]开发了HELM框架,实现了将HDL(硬件描述语言)程序自动转换为HE电路的功能,支持多种密文评估模式。与此同时,文献[151]提出了一个自动化自举管理编译器DaCapo,一定程度上降低了手动规模管理与自举放置带来的复杂性与性能开销,有望进一步提升CCNN推理与训练效率。

3.2.3 面向大规模高效推理的FHE-NN

针对HE在大规模数据处理中的性能瓶颈,文献[152]提出了HE加速器CHAM,采用Xilinx FPGA实现了高效的矩阵-向量乘积加速,支持多种HE操作类型及其转换,以期提升端到端应用性能。文献[153]指出FHE在大规模部署用于安全神经推理时面临计算成本高、将神经网络有效映射到FHE原语困难,以及编程方面如大向量打包、噪声管理、程序转换等诸多挑战,使得用现有工具构建大型FHE神经网络难以实现。文中提出了全自动化使用FHE进行CCNN推理的框架Orion,该框架可接受用PyTorch编写的DNN,并将其转换为高效的FHE程序。同时,Orion框架还提出了针对任意卷积的单次多路复用打包策略,以及一种新的、高效的自动化自举放置和规模管理技术,但是,该打包策略在面对其他应用场景时的扩展适应能力,还需进一步提高,同时,对计算资源的占用问题仍待进一步研究。

为了进一步提升基于FHE的NN的训练和推理效率,文献[154-155]通过分布式高性能计算技术,通过多服务器多处理器的流水线技术,推动CCNN的并行化计算。文献[156]针对FHE无法直接评估CNN中非算术激活函数的问题,以及现有方法在精度损失和延迟增加之间需权衡的挑战,提出了LPFHE框架。LPFHE框架能够使用低复杂度多项式精确逼近CNN中关键的ReLU函数,且支持为每个ReLU函数找到最优的逼近域和多项式。通过将分段加权最小二乘法算法与Remez[157]算法相结合,LPFHE相较于实现了更高的逼近精度。LPFHE能够生成具有高推理精度的低复杂度多项式CNN,因为低阶多项式能很好地保留ReLU函数的特性。

同态加密因具备数学可验证的安全性,且支持单方计算,受到广泛关注。然而,实验结果(见文献[158]表2与文献[159]表5)表明,密文膨胀已成为其主要瓶颈,不仅导致通信量显著增加,还引入了高开销的自举操作,限制了其在实际系统中的应用。

3.3 基于ZKP技术的NN计算证明

ZKP是一种重要的密码学技术,它允许一方(证明者)在不泄露任何除结论本身之外的额外知识的前提下,向另一方(验证者)证明其陈述的真实性[160-161]。这种特性使得ZKP在保障数据隐私和增强信息安全方面具有极大潜力[162]。随着研究的不断深入,ZKP技术经历了从理论提出到高效实用化的持续演进。

1985年,零知识证明及交互式零知识(Interactive Zero Knowledge, IZK)概念由Goldwasser等人首次提出[163],并对证明系统进行了形式化定义。此后,经过多年的理论积累与技术创新,ZKP迎来了实用化的转折点。文献[164]中,Bitansky等人提出了零知识简洁非交互式知识论证(Zero Knowledge Succinct Non-Interactive Argument of Knowledge,zkSNARK)模型,使得验证过程可在短时间内完成,且证明尺寸极小,仅需数个字节。同时,zkSNARK实现了验证者离线验证、证明者多项式时间生成等特性,极大提升了应用可行性。

在zkSNARK的基础上,研究者们持续推动理论与应用的发展。文献[164]提出了构建预处理简洁非交互式论证(preprocessing SNARG)的通用方法,进一步指明了研究趋势正从单纯优化论证长度,转向关注验证时间最小化与计算委托问题。与此同时,Ben-Sasson等人在文献[165]中提出了基于交互式预言机证明(interactive oracle proof, IOP)协议和里德-所罗门码接近性测试的零知识可扩展透明知识论证(Zero-Knowledge Scalable Transparent Argument of Knowledge,zkSTARK)系统,提供了一种无需可信设置、抵抗量子攻击的新型ZKP方案。

随着ZKP技术基础的逐渐夯实,其在DL推理和神经网络应用领域迅速扩展。文献[166]基于zkSTARKs提出了用于高效可验证神经网络推理的优化编译器,推动了ZKP在AI推理中实际落地的后续研究。进一步地,文献[167]提出了首个专门针对LLM的ZKP系统zkLLM,以满足验证LLM输出

真实性的立法需求。此项工作不仅提出了适用于非算术张量运算的并行化查找论证方法tlookup，还开发了专用零知识证明zkAttn，加速了注意力机制的验证过程，并引入GPU并行化加速推理。

同时，针对CNN推理的验证问题，文献[168]基于zkSNARK提出了vCNN框架，设计了新型卷积关系表示方法，将证明复杂度由O(ln)降至O(l+n)，为进一步研究推理证明的效率与资源消耗的平衡问题提供了经验。

在NN模型层面，文献[169]提出了ZKML框架，通过从TensorFlow到halo2电路的优化编译器，支持包括视觉模型、蒸馏版GPT-2在内的先进模型的zkSNARKs生成。针对百万级电路规模和标量级依赖问题，文献[170]提出了ZENO优化框架，引入保留高级语义与隐私、张量信息的ZENO语言结构，并设计了隐私类型和张量类型驱动的多层次优化方法，大幅提升了神经网络推理时ZKP系统的效率。然而，尽管诸多研究工作在证明生成方面取得了很大进展，但生成证明和进行验证的资源消耗和计算速度问题仍然悬而未决。

进一步地，文献[171]关注于提升GPU加速系统的吞吐量，提出了ZKP的全流水线式批量生成系统，为基于ZKP的神经网络推理加速开辟了新方向。

围绕DL应用中用户验证需求的实际问题，文献[172]提出了zkCNN方案，允许模型所有者在不泄露模型参数的情况下，向他人证明预测结果确实由指定CNN模型计算得出，并扩展到CNN在公开数据集上的验证。同时，该工作在傅里叶变换、卷积验证协议上进行了创新，实现了证明者时间的线性扩展。

在DL训练阶段，文献[173]提出了Kaizen框架，针对DNN训练提出了具有可证明安全性、隐私保护、简洁证明和高效验证的零知识训练证明(Zero-Knowledge Proof of Training，zkPoT)方法，支持动态小批量梯度下降而无需固定迭代次数，为灵活训练提供了保障。

同样针对DL训练过程中的隐私验证问题，文献[63]提出了zkDL方法，特别设计了针对ReLU激活函数的专用证明zkReLU，并开发了FAC4DNN电路结构，通过聚合跨层、跨步骤的证明，支持CUDA高效实现，极大提高了DL训练过程的验证效率。

针对大规模CNN训练完整性验证问题，文献[174]提出了VeriCNN方案，结合Winograd算法优化卷积操作与高概率矩阵乘法，提高了在大规模数据处理下的验证效率与证明性能。

为了进一步拓展ZKP在大规模计算上的适用性，文献[175]引入了Hekaton zkSNARK方案，通过“分布-聚合”框架，支持大型计算任务的并行分块证明与简洁聚合，有效解决了传统zkSNARK在处理超大计算量时的性能瓶颈。

从更广阔的角度来看，zkSNARK与zkSTARK两种主流技术体系在设计理念和安全假设上各具特色。与SNARKs依赖于复杂的数学假设不同，STARKs的现有主流构造通常基于抗碰撞哈希函数，从而具备后量子安全性优势[176-177]，同时也避免了可信设置[178]，并具备高度的通用性。尽管zkSTARK尚处于持续发展阶段，但其在大规模神经网络训练与推理中的潜力正被广泛期待。

而在面向数据并行算术电路的ZKP中，文献[179]提出了空间高效的Sparrow方案，证明者空间开销以小于对数因子的速率增长，且空间需求渐近小于电路规模，为CC下基于ZKP的神经网络训练与推理提供了强有力的支撑，但Sparrow仅支持数据并行电路。

近年来，基于ZKP的FL方向也取得了显著进展[159]。文献[180]提出了zPROBE框架，基于ZKP推导统计界限，有效实现了对恶意更新的检测与清除，增强了FL系统的拜占庭容错能力。文献[181]则提出了基于零知识联邦学习(ZKFL)的新型全局模型聚合方法，利用Chord覆盖网络实现了即便在恶意环境下的聚合可信性保障。

此外，文献[182-186]等也提出了多种可扩展、可审计、可验证的联邦学习方案，为基于CC与ZKP的机器学习研究提供了丰富的工具和思路。

最后，ZKP技术在节省计算资源、抵御恶意攻击等方面也展现出巨大潜力[187-189]。例如，文献[190]针对智能网联汽车(Intelligent Connected Vehicles，ICV)环境下的数据利用问题，提出了基于区块链和zkSNARKs的BV-ICVs联邦学习框架，通过智能合约化验证，有效防止不可靠模型更新，能够极大增强系统安全性。

本节深入分析了当前基于密码学原语的CCNN研究工作的技术基础及优点，并对一些研究工作的未来方向进行了梳理，其中的优秀研究成果如表3所列。

表 3 基于密码学原语的优秀CCNN研究一览

文献	基础技术						应用		计算开销	通信开销	精度损失	安全目标	敌手模型
	SS	**GC**	**HE**	**OT**	**ZKP**	**DP**	NN	FL					
[90]	√		√		√				高	高	中	可验证性	主动敌手
[91]				√			√		高	中	高	计算正确性	主动敌手
[95]	√	√		√			√		高	高	中	数据机密性	被动敌手
[99]		√		√			√		中	中	高	计算正确性	主动敌手
[100]	√	√	√	√			√		高	高	中	计算正确性	主动敌手
[104]	√						√		高	中	中	数据机密性	被动敌手
[106]	√	√		√			√		中	高	中	数据机密性	被动敌手
[109][110][111]	√						√		高	高	中	数据机密性	被动敌手
[112][113]	√	√		√			√		高	高	中	数据机密性	被动敌手
[115][116][117]	√		√				√		中	高	中	数据机密性	被动敌手
[126][127]			√				√		高	高	低	计算正确性	主动敌手
[128][129][130]			√				√		高	高	中	计算正确性	主动敌手
[131][132]			√				√	√	高	高	低	计算正确性	主动敌手
[133][134][135][136] [137][138][139]			√				√	√	高	中	低	数据机密性	主动敌手
[140][146]			√				√	√	高	中	低	计算正确性	主动敌手
[147][148][150][151]			√				√	√	高	高	中	数据机密性	被动敌手
[152][153][154][155][156]							√	√	高	高	中	计算正确性	主动敌手
[164][165]					√				高	高	-	可验证性	被动敌手
[165]					√		√		高	高	高	可验证性	主动敌手
[166][179]					√		√		高	高	中	可验证性	被动敌手
[167]					√		√		高	高	中	可验证性	被动敌手
[168][169][170]					√		√		中	高	中	可验证性	被动敌手
[171]					√		√		高	中	中	可验证性	被动敌手
[63][172][173][174][175]					√		√		高	高	中	可验证性	被动敌手
[180][182][183][184]					√			√	高	高	中	可验证性	被动敌手
[189][190]					√		√		中	高	中	可验证性	被动敌手

注：本表中“高 / 中 / 低”分别依据如下标准评估：计算开销指相对明文模型计算时间 ＞100 倍为“高”，10-100 倍为“中”，＜10 倍为“低”；通信开销以数据量与交互频率判断；精度损失按准确率下降幅度划分。详情参考第1节。

4 基于混合架构的CCNN

本文所称的“混合架构”是指在同一私有推理或训练流程中，综合利用 HE、SMPC、ZKP 等不同安全计算技术分别支撑神经网络的不同计算阶段的系统。在此类系统中，不同类别的密码原语通常在算子边界处进行密文格式或域间转换，该机制已成为协议设计的标准组成部分，而非仅作为性能优化的特例。相较于纯 HE 或纯 SMPC 等单一范式的框架，混合架构能够在相同或相近的安全模型下，实现更优的计算吞吐与通信效率平衡。通过在体系中有机融合 HE、FL 与 SMPC 等隐私计算技术，该类架构进一步拓展了混合框架的设计理念，不仅在性能与安全性之间取得更优平衡，还能够支持跨参与方的数据协同与端到端的隐私保护训练与推理[191-195]。

这一领域的理论基础可以追溯至1972年文献[196]提出的公钥密码体制，为公钥加密在隐私计算中的应用奠定了基础。1982年，姚期智教授提出了安全多方计算的概念，并通过“百万富翁问题”[197]，开启了CC领域的新篇章。随后，1985年，Goldwasser、Micali和Rackoff在文献[163]中首次提出ZKP的概念，进一步丰富了隐私计算的密码学工具体系。上述理论成果为后续混合架构中安全计算和隐私验证机制的发展提供了坚实支撑。

随后，研究者们进一步探索了混合加密与2PC结合的高效推理系统。文献[28]提出Gazelle系统，设计了高效的HE库与线性代数内核，并通过加密切换协议无缝衔接HE与GC，从而实现了低延迟推理。进一步地，CrypTFlow2[198]与Delphi[118]分别在

OT协议基础上，提出了适用于深度NN推理的高效两方加密框架。Cheetah框架[119]则针对两方神经网络（Two-Party Computation Neural Network）推理系统高开销问题[199-201]，特别是在处理ResNet50等大型网络时，提出了多种轻量化、通信高效的非线性函数原语，进一步提升了两方推理系统的整体性能。基于此思路，文献[199]开发了Panther框架，该框架通过定制HE算法以减少多项式乘法，并优化了百万富翁协议，降低了通信与计算开销。为了进一步提升推理效率和带宽利用率，文献[202]基于SS提出了FPCNN方案，采用缩放噪声模型与安全并行ReLU协议，有效优化了CNN的线性与非线性层操作，而文献[203]则提出Swift系统，通过结合FHE与SS，在保障隐私的同时，加速了ReLU与最大池化等非线性计算，建立了高效的编码转换机制。但是为了进行不同协议之间的相互转换以及在SS下执行计算，这两个工作的通信负载仍然有待进一步优化。

为突破两方计算在恶意威胁模型下的局限性，文献[204]基于计算密集型的分层HE[90]和通信密集型的SS Beaver三元组[205]协议，将2PC协议MUSE[206]和SIMC[207]推广到三方计算（Three-Party Computation，3PC）环境，提出了AUXILIATOR和SOCIUM协议，有效提升了安全性与性能，但其通信负载和计算开销仍然太高。

与此同时，结合DP与SMPC的研究工作也受到关注。文献[208]提出高效的离散高斯分布采样方案，兼顾了通用性与高效性，避免了昂贵的非整数计算。文献[209]则在抵御拜占庭攻击者的前提下，扩展了分布式噪声生成协议，并对不同离散采样协议进行了系统评估。进一步地，在新型密码系统的设计方面，文献[210]提出了基于分布式双陷门公钥密码系统（Distributed Dual Trapdoor Public-Key Cryptosystem，DT-PKC）的隐私保护CNN分类方案，结合了安全乘法与Tanhplus函数优化ReLU激活计算。文献[211]则从计算调度角度出发，提出了Seesaw框架，通过减少非线性操作并重复使用线性计算结果，在确保精度的同时加速了NN推理。这几项工作基于HE和SMPC（包括SS、DP等）以及其他技术，实现了非线性操作的执行，但计算和通信开销仍然不具备现实部署的条件，但为完善CCNN训练和推理技术生态提供了研究思路和方向。

为了进一步释放硬件潜力，一些研究工作开始结合GPU优化、编译技术及数据打包策略来提升CCNN的整体训练与推理效率[212]。例如，文献[213]提出BitPacker，通过固定大小字打包方式，提升CKKS在加速器上的运算效率。文献[214]则提出Rhombus协议，基于环上容错学习（Ring Learning With Errors，RLWE）与带系数编码HE，支持更快速的非线性计算。文献[215]提出了基于块循环变换的高效私有推理框架PrivCirNet，通过将网络权重结构化并在频域中实现加密计算，显著降低了同态加密与多方安全计算的通信与计算开销。该方法在保持模型精度的同时统一了不同隐私保护范式下的推理接口，展示了结构化矩阵与频域变换在提升私有推理效率与可扩展性方面的潜力，为高性能隐私保护深度学习提供了新的研究方向。

针对当前最广泛应用的Transformer模型，也涌现出如文献[105，108，216-220]等多项基于机密计算的研究成果。其中，文献[219]提出了一系列针对Transformer关键操作（如矩阵乘法、Softmax、GeLU和LayerNorm）的高效安全协议，结合定制化矩阵乘法与紧凑打包技术，提升了整体推理效率。面向自然语言生成（NLG）场景，文献[220]开发了MERGE框架，通过重用输出隐藏状态、重组Transformer模块线性操作，有效加速了基于Transformer的隐私文本生成推理过程。文献[108]基于SecureNLP[105]提出SecureGPT框架，这是一种适用于GPT的多方隐私保护Transformer推理框架，该框架中包含一系列构建模块，包括乘法份额到加法份额的转换（M2A）、截断、除法、激活函数Softmax和GeLU协议，同时为GPT的Transformer子层设计了多方隐私协议，但该框架在大型网络上仍然存在通信负载高的问题。

文献[221]提出了面向Transformer模型的高效安全计算框架BumbleBee，通过结合同态加密与多方安全计算技术，设计了针对矩阵运算的优化协议，实现了在不修改模型结构的情况下高效支持多种预训练模型的隐私保护推理。该框架在通信效率与计算开销方面均较现有方案有显著改进，并展现了良好的工程可复现性与系统扩展性。随后，文献[222]在此基础上提出了混合安全推理框架BLB（Breaking Layer Barriers），首次突破传统层级化推理的限制，通过线性算子的细粒度融合与安全的CKKS－SMPC转换机制，有效降低了模式切换带来的通信开销，并进一步优化了多头注意力等关键算子的安全执行效率。但是，计算复杂度、通信负载

和模型准确率下降，仍然是针对基于Transformer的CCNN需要解决的关键问题。

综合而言，基于混合架构的CCNN技术，正在从早期的基础密码学原理支撑，逐步发展到专门针对二进制NN、CNN乃至Transformer等不同模型架构的高效安全推理方案，并不断结合硬件优化、协议与网络协同设计，形成一条技术持续演进与深化的清晰脉络。表4对基于混合架构的优秀CCNN研究进行了总结，可以看出，现有基于混合架构的CCNN研究整合了各种底层技术的优点，兼顾了一些CCNN性能和机密性要求，为以后的研究工作提供了经验。

表4 基于混合框架的优秀CCNN研究一览

文献	基础技术						应用		计算开销	通信开销	精度损失	安全目标	敌手模型
	SS	GC	HE	OT	ZKP	DP	NN	FL					
[118,119,198]	√						√		高	高	中	数据机密性	被动敌手
[120,121]		√		√			√		高	高	中	数据机密性	被动敌手
[122]	√			√			√		高	中	中	数据机密性	被动敌手
[25,204]	√		√				√		高	高	中	计算正确性	被动敌手
[28,199]		√	√				√		高	高	中	计算正确性	主动敌手
[202,203,215]	√		√				√		高	高	中	数据机密性	被动敌手
[208-210]	√		√			√	√		中	高	中	数据机密性	被动敌手
[213,214]			√				√		中	高	中	计算正确性	主动敌手
[105,108,219-222]	√		√				√	√	高	高	中	数据机密性	被动敌手

注：本表中“高/中/低”分别依据如下标准评估：计算开销指相对明文模型计算时间＞100倍为“高”，10-100倍为“中”，＜10倍为“低”；通信开销以数据量与交互频率判断；精度损失按准确率下降幅度划分。详情参考第1节。

5 硬件支持下的系统优化

随着CC需求的不断增长，研究者们尝试从不同角度探索神经网络领域中的隐私保护问题。在此过程中，存内计算（Compute-in-Memory，CiM）架构逐渐引起了广泛关注。文献[223]基于布尔可满足性（Boolean Satisfiability Theory，SAT）理论[224-226]，提出了一种专门针对存内计算架构的去混淆方法CiMSAT。作者针对CiM的存储和混合信号计算特性进行了深入安全分析，创新性地拟合并利用了“无推断值”的混淆数据以进行函数逼近，同时设计了偏差容忍的布尔可满足性算法，以修正由于逼近产生的偏差并成功重构混合信号电路。这一研究为存内计算架构在隐私保护方面的应用及其与其他隐私计算技术的混合架构设计提供了有益参考[227-228]。

在存内计算领域取得进展的同时，可扩展神经网络计算中的通信开销问题也成为瓶颈。为了解决这一挑战，文献[229]提出了双重拉格朗日编码（Double Lagrange Coding）机制。这种新型迭代编码方法使得多方多项式迭代计算具备线性通信复杂度，从而在整个迭代过程中保持并行化优势、对抗容忍度以及节点退出后的恢复能力。相比传统的二次通信开销，该方案实现了显著优化，为大规模隐私神经网络计算提供了新思路。

随着对高效计算资源的需求不断提升，研究者们也开始将目光转向GPU（Graphics Processing Unit）与FPGA（Field-Programmable Gate Array）[230]等硬件加速器。在加速器领域，针对多标量乘法（Multi-Scalar Multiplication，MSM），涌现出了一系列创新设计，同时，MSM也是ZKP生成过程中的核心操作。

首先，在GPU加速方面，文献[231]提出了一种专为分布式多GPU系统定制的新型多标量乘法算法DistMSM。该算法对经典Pippenger算法[232]进行了调整优化，使其适应多GPU架构特点。具体而言，DistMSM在GPU内核层面引入了针对当代GPU架构定制的椭圆曲线算术内核，并通过两种创新技术有效缓解寄存器压力，同时利用张量核心（Tensor Cores）加速特定的大整数乘法运算。

进一步地，在可重构硬件加速方面，文献[233]提出了高度可重构的加速器ReZK，旨在加速ZKP生成阶段的关键计算，尤其是数论变换（NTT）与逆数论变换（INTT）[234]、以及多标量乘法（MSM）[235]。ReZK通过灵活调整片上存储器与运算核心之间的数据路径，可配置为不同规模和位宽的NTT、INTT或MSM模式，展现了极高的适应性与性能

潜力。

在FPGA平台上，针对MSM任务的专用加速器设计也取得了重要进展。文献[236]提出了MSMAC，一种用于大规模多标量乘法的FPGA加速器。MSMAC引入了专门为MSM设计的指令集架构（Instruction Set Architecture，ISA），并通过混合型卡拉苏巴乘法器（Karatsuba Multiplier）[237]优化了流水线点加法单元（Point Addition Unit，PAU）。此外，作者设计了一个高效的运行时系统，可以根据任务规模最优地划分子任务，并协调多个处理单元（Processing Elements，PEs）协同执行，从而大幅提升了整体计算效率。

与此同时，文献[238]提出了另一款针对MSM运算的FPGA加速器Falic。考虑到MSM是zkSNARK（零知识简洁非交互式知识论证）证明生成过程中最耗时的环节，Falic在架构设计上作出了三项关键贡献：首先，采用全局异步局部同步（Globally Asynchronous Locally Synchronous，GALS）策略，构建了多个小型轻量级MSM内核，实现标量向量与点向量不同部分的独立并行化内积计算；其次，每个内核内仅包含一个大整数模乘器（Large-Integer Modular Multiplier，LIMM），通过复用方式高效完成点加法（PADDs）操作；最后，通过简单的缓存结构实现了计算复用，从而进一步提升了资源利用率和加速效果。

综上所述，从存内计算方法到通信高效的编码机制，再到GPU和FPGA平台上的硬件加速设计，这些研究体现了CCNN领域技术演进的多元方向，也为未来CCNN设计与优化提供了丰富参考路径。

6　CCNN系统应用

随着数据隐私保护与人工智能融合的趋势日益加深，CCNN逐渐成为构建安全可信AI系统的核心支撑技术之一。通过在受保护的计算环境中执行模型训练与推理，CCNN能够在保障数据机密性的同时实现高效的智能决策。这一特性使其在金融风控、医疗影像分析、工业制造优化等高敏感场景中展现出独特优势。本节将系统性地介绍CCNN的典型系统框架与应用方向，并通过比较分析不同技术路线的特点与适用场景，为后续研究提供技术选型参考。

6.1　现有可用的CCNN系统框架介绍

为了支撑CCNN的高效开发与部署，业界与学术界陆续提出了多种系统实现框架。这些框架在底层安全机制、计算范式以及开发抽象层次上各具特点，形成了丰富的技术生态。从TEE的轻量化框架，到依托SMPC和HE的分布式系统，不同方案在性能、安全性与可用性之间做出了不同权衡。本小节将梳理当前主流的CCNN系统框架与相关开源实现，分析其设计理念、功能特征与典型应用场景（如表5所示），为后续的比较与选型分析奠定基础。

最早期的开源探索之一是文献[239]提出的PySyft框架。PySyft为CCNN提供了重要工具，集成了HE、SMPC和DP[240]三种机密计算技术，并创新性地采用了基于命令链和张量的表示方法。这一方法支持FL、SMPC和DP等复杂隐私保护机制，同时保留了终端用户熟悉的深度学习API。PySyft由Pyffel等人最早提出，其第一个版本由去中心化AI平台OpenMined主导开发。

随后，微众银行（WeBank）AI部门于2019年推出了第一个开源工业级FL框架FATE[241]。FATE采用了HE和SMPC两种机密计算技术，支持多个企业在不暴露数据的前提下进行协作式机器学习。同年，谷歌开发了TensorFlow Federated（TFF）[242]，这是一个专门面向FL和多源数据联合计算的开源框架。TFF不仅使开发者能够在本地模拟分散式计算环境，还允许对联邦算法进行实验和优化，并支持在多源数据上进行聚合分析等非学习计算。

进入2020年，针对PyTorch生态，文献[243]提出了FL框架FedML。FedML是一个开放的研究库和基准平台，旨在促进FL算法的开发与性能公平比较，支持设备端训练、分布式计算和单机模拟

三种计算范式。与此同时，文献[244]提出了Flower框架，文献[245]提出了模块化基准测试框架LEAF，均为FL领域提供了不同层面的支持和补充。

针对联邦环境下数据异质性问题，文献[246]提出了FederatedScope平台。FederatedScope采用事件驱动架构，为用户描述不同参与者的行为提供了极大的灵活性，能够支持具有不同本地训练过程、学习目标和后端配置的参与者，并协调它们采用同步或异步策略进行联邦训练。

除了上述框架，业界也相继推出了多个重要的开源平台，如华为公司的MindSpore Federated框架[247]、字节跳动公司的Fedlearner框架[248]、百度公司的PaddleFL框架[249-250]、矩阵元技术公司的

表5 主流CCNN系统框架概览

框架名称	提出机构/作者	支持技术	支持能力	典型特色
PySyft	OpenMined等	HE,SMPC,DP	联邦学习、张量抽象	命令链表示法,支持多种隐私机制,兼容主流深度学习API
FATE	微众银行	HE,SMPC	企业间协作训练	工业级应用框架,支持水平/垂直建模,适配金融等场景
TensorFlow Federated (TFF)	Google	FL	联邦算法实验与模拟	多源数据聚合支持,本地开发环境友好
FedML	He等	FL	多范式计算训练与评估	支持设备端、单机、分布式三类训练范式,PyTorch生态兼容
Flower	Beutel等	FL	客户端-服务器协同	轻量高效,接口简洁,异构终端兼容性强
LEAF	Caldas等	FL	数据集划分、基准测试	模块化设计,便于公平性评估与算法对比
FederatedScope	Xie等	FL	异质性支持、事件驱动	支持异步/同步策略,参与者行为高度可定制
MindSpore Federated	华为	FL	专家调度、大模型训练	支持EfficientMoE调度机制,国产平台适配性好
Fedlearner	字节跳动	FL	多方隔离训练协作	强调隐私边界与合规性,企业协作友好
PaddleFL	百度	FL	分布式隐私学习集	集成于PaddlePaddle,适配国产芯片
Rosetta	矩阵元技术	FL + TEE	企业部署、可信执行	强化安全机制,强调实际可用性
CoMind	Zhu等	TEE	GPU加速执行	与TFF接口兼容,支持GPU加速
CrypTen	Facebook	SMPC	PyTorch集成、加密训练	面向研究用户,适配性与易用性强
FALCON	Wagh等	SMPC	多方安全计算	半诚实安全模型,适配神经网络训练
FedLLM-Bench	Ye等	FedML	大语言模型联邦训练	涵盖8种训练方法与6项评估指标,测试体系完备
ClaraFL(闭源)	NVIDIA	TEE/FL	医疗图像、推理任务	面向医疗场景,强调能效与模型隐私
FMPC(闭源)	Garg等	FL + TEE	跨组织协作	多组织可信协同,强调可扩展性

Rosetta框架[251]。此外,还有如CoMind框架[252]、CrypTen框架[95]、FALCON框架[112]等。值得注意的是,TFF与CoMind框架支持在GPU上执行,充分利用硬件算力,加速机密计算环境下的训练与推理过程。与此同时,一些闭源框架如ClaraFL[253]和FMPC[37]也在探索更高效的机密计算AI应用。

在这些基础设施的推动下,研究者们不断针对新兴需求进行拓展与深化开发。文献[254]在FedML框架基础上提出了FedLLM-Bench框架,该平台涵盖了8种训练方法、4个训练数据集以及6种评估指标,极大地丰富了联邦大语言模型研究的测试环境。另一方面,文献[255]提出了EfficientMoE方法,针对专家负载和数据特征的调度进行了优化,并在MindSpore平台上实现,进一步扩展了MindSpore在联邦学习和大模型训练场景下的适用性和灵活性。

综合而言,基于机密计算的AI系统应用正经历从早期功能性框架建设,向面向特定场景优化与性能强化方向不断演进。随着框架体系日益成熟,未来机密AI系统将在隐私保护、算力利用与跨组织协作等方面发挥更大的潜力。

6.2 系统比较与技术选型分析

在多种CC框架并存的背景下,技术路线的合理选择成为系统设计与落地的关键问题。不同安全计算方案在信任模型、性能开销和开发复杂度上各有优势与局限。表6对主流技术路线进行了系统比较:TEE依赖可信硬件环境,性能接近原生但受限于I/O与内存;SMPC具备较强安全性和协作能力,却存在通信延迟;HE可在单方环境下实现非交互式计算,但计算复杂度较高;ZKML则以可验证性为特征,适用于高安全推理任务。进一步地,表7总结了典型应用场景下的选型建议:TEE适合实时性要求高的场景,SMPC与HE更适用于多方隐私协作,而ZKML则面向结果可验证与合规需求的应用。该比较框架可以为后续CCNN系统的技术选型与实践提供清晰参考。

表7总结了各类应用场景的主流方案,其性能优势主要源于底层算法与系统机制的差异化设计。对于云端私有推理任务,当前SOTA方案多采用基于CKKS的同态加密推理框架,通过密文批处理与多项式近似降低乘法深度,并在GPU TEE(如NVIDIA Hopper)中实现加速与远程证明,从而在隐私保护与性能之间实现高效平衡。在多方联合训练场景中,主流方案通常结合安全聚合机制与轻量级SMPC或HE+SMPC混合结构,利用Beaver三元组和OT协议实现高效梯度聚合,并辅以差分隐私技术控制模型更新过程中的潜在泄露。边缘端执

表6　不同机密计算与密码学技术路线的系统比较

比较维度	硬件TEE	SMPC	HE	ZKML
安全模型与信任假设	依赖硬件厂商信任；可能受侧信道影响	数学安全性，需大多数参与方诚实	数学安全性，单方计算即可	命令链表示法，可验证计算，防止恶意推理结果
性能开销	接近原生速度，但I/O昂贵	通信频繁，延迟高	无交互，但计算量巨大	生成证明耗时长，验证较快
可扩展性	易部署，单机或集群	扩展性受通信限制	计算可并行化，支持大模型加速	可扩展性低，主要用于推理验证
易用性与改造成本	代码改动小，可复用原生框架	需重写协议或函数级安全算子	需支持特定的密文算子	需转换为电路或算子约束形式
适用阶段	训练、推理均可	多方协作训练/推理	单方私有数据训练或推理	推理结果公开验证
典型代表	Intel SGX, AMD SEV, NVIDIA Hopper	ABY, CrypTFlow, SecureNN	CKKS, BFV, TenSEAL	zkML, zkSNARK, ezkl
主要瓶颈	EPC大小、I/O频繁	网络延迟、带宽负载	计算复杂度、乘法深度、近似误差	证明时间、内存占用

行则主要依托轻量级TEE（如TrustZone、SGX）运行剪枝或量化后的模型，通过常数时间算子与远程证明机制降低侧信道攻击风险。针对模型验证与可证明推理任务，ZKML与zkSNARK技术逐渐成为主流，其核心在于将神经网络算子电路化，并通过多项式承诺与查找表优化降低证明生成开销，从而在保障推理正确性的同时保持系统可扩展性。在跨域医疗与金融协作场景中，SMPC与HE的混合方案兼具合规性与可扩展性：通常先通过私有集合求交实现样本对齐，再结合HE完成密态统计与梯度更新。最后，在大模型训练与推理加速场景中，GPU TEE或结合分层SMPC的混合框架能够在保持隔离性与安全性的同时，利用安全通信与内存加密支撑高性能分布式训练。

表7　不同应用场景下的机密计算技术选型建议

应用场景	典型需求	推荐方案	原因
云端私有推理（单方输入）	数据保密＋模型完整性	HE或GPU TEE	HE无需交互；GPU TEE原生加速
多方联合训练（横向/纵向）	多数据方输入＋联合梯度	SMPC或混合框架（HE＋SMPC）	SMPC支持交互式聚合；混合框架兼顾性能
边缘端安全执行	小模型＋设备可控	TEE（TrustZone，SGX）	轻量、低延迟，硬件支持广
模型验证（可证明推理）	输出可验证＋第三方信任	ZKML或zkSNARK	提供可验证性与零知识性
跨域医疗/金融协作	法规严格＋高隐私	SMPC＋HE混合	安全模型强、可扩展性平衡
大模型训练加速	高性能＋数据隔离	GPU TEE（Hopper）	支持HBM加密和安全通信

7　总结与展望

如前文所述，将这些技术应用于神经网络训练与推理，需应对一系列结构性挑战。例如，NN中的大量线性算子（如矩阵乘、卷积）虽结构规则但资源密集，如何在密态中高效实现这些算子是系统性能的核心瓶颈；而非线性函数（如ReLU、Sigmoid、Softmax）在受保护环境中难以直接计算，常需通过多项式逼近或协议转换实现；此外，隐私保护下的梯度反向传播、参数更新以及优化算法设计，也对通信同步、带宽消耗与精度控制提出了更高要求。

尽管CCNN的训练与推理技术已取得显著进展，在数据机密性与完整性保护方面展现出独特优势，但在实现高效、可扩展的模型训练与推理系统方面仍存在诸多挑战，CCNN训练和推理工作仍需进一步研究。

首先，针对模型复杂性支持不足的问题，开发密文友好算子，如以低阶多项式、线性近似或低通信复杂度为原则重构激活函数、归一化操作、注意力机制等，将有助于提升硬件计算资源等对复杂计算的支持和在受限Enclave环境高效执行机密计算操作。同时，需要在资源约束条件（如内存限制、计算图静态化需求）下，发展模块化、浅层但高效的机密计算网络架构。此外，需要建立跨层高效内存管理机制，即针对Enclave环境内存容量有限、HE下密文膨胀严重的问题，开发跨层数据流优化策略，包括中间状态压缩、分层缓存管理、智能分页和按需加载技术。

其次，针对系统可扩展性与应用性不足的问题，需要进一步研究隐私感知的负载调度策略，根据密文任务的特性（如计算量、通信需求）设计调度算法，实现资源高效利用且私密可靠的数据利用。同时，需探索弹性扩展的加密通信协议，适配动态变化的节点数量和网络条件，保障系统在弹性伸缩过程中的安全一致性。

此外，从不同技术路径的角度来看，基于硬件TEE的方案具有较高的原生执行效率与系统透明性，适合部署在受控环境下进行高性能推理任务；而基于密码学原语的方法（如SMPC、HE、ZKP）则具备更强的理论安全性与灵活性，适用于跨域数据协作、强安全约束或无可信硬件支持的环境。其中，HE更适用于低交互、模型已定型的单方推理任务，而SMPC更适合高交互、高通信带宽的联合训练任务。ZKP则逐步在可验证训练、合规审计等场景中显示出优势。

当前，针对典型的隐私计算应用场景，已有多项技术实现达到实用化水平：如联邦学习系统中的FATE、FedML、MindSpore Federated；用于神经网络隐私训练的SecureTF（TEE）、MP-SPDZ（SMPC）、Kaizen（ZKP）；用于私密推理的Cheetah、Delphi、AutoFHE等。这些系统分别代表了各类技术范式下的可部署路径，为CCNN的工程落地和研究探索提供了坚实基础。

总体而言，尽管当前技术尚处于初步阶段，存在诸多挑战，但随着硬件技术演进（如可扩展TEE架构、内存加密优化）、高效密码协议设计（如轻量级SMPC、近似HE方案）以及系统软件生态的发展，CCNN应用将在保障数据隐私与释放AI潜力之间实现更优的平衡。期望通过本综述总结与分析，为后续在不同应用场景中选择合适的CCNN技术路线提供明确的参考依据。

致 谢 感谢《计算机学报》编辑和审稿专家，他们付出了辛勤的工作。

参 考 文 献

[1] Fouda Mostafa M., Fadlullah Zubair Md, Ibrahem Mohamed I., Nei Kato. Privacy-preserving data-driven learning models for emerging communication ion networks: A comprehensive survey. IEEE Communications Surveys &. Tutorials, 2025, 27(4):2505--2542

[2] Zhang He, Wu Bang, Yuan Xingliang, Pan Shirui, Tong Hanghang, Pei Jian. Trustworthy graph neural networks: Aspects, methods, and trends. Proceedings of the IEEE, 2024, 112(2):97-139

[3] Chattopadhyay Arup Kumar, Saha Sanchita, Nag Amitava, Nandi Sukumar. Secret sharing: A comprehensive survey, taxonomy and applications. Computer Science Review, 2024, 51:100608-100630

[4] Divya B., Dhas Anto Sahaya. A systematic literature review on dynamic load balancing techniques in cloud computing environment: Techniques, challenges, and future prespects// Proceedings of the 2nd International Conference on Machine Learning and Autonomous Systems (ICMLAS 2025). Kanyakumari. Tamil Nadu, India, 2025: 1440-1446

[5] Qayyum Adnan, Qadir Junaid, Bilal Muhammad, Al-Fuqaha Ala. Secure and robust machine learning for healthcare: A survey. IEEE Reviews in Biomedical Engineering, 2020, 14: 156-180

[6] Jonnagaddala Jitendra, Wong Zoie Shui-Yee. Privacy preserving strategies for electronic health records in the era of large language models. npj Digital Medicine, 2025, 8(1):34

[7] Javaheripi Mojan, Chen Huili, Koushanfar Farinaz. Unified architectural support for secure and robust deep learning// Proceedings of the 57th ACM/IEEE Design Automation Conference. San Francisco, USA, 2020: 1-6

[8] Andreoli R., Mini R., et al. A multi-domain survey on time-criticality in cloud computing. IEEE Transactions on Services Computing, 2025, 18(2):1152-1170

[9] Dhasarathan Chandramohan, Hasan Mohammad Kamrul, et al. User privacy prevention model using supervised federated learning-based block chain approach for internet of medical things. CAAI Transactions on Intelligence Technology, 2023, 8(4):1097-1112

[10] Ettaloui Nehal, Arezki Sara, Gadi Taoufiq. An overview of blockchain-based electronic health record and complianc e with gdpr and hipaa. Data and Metadata, 2023, 2:405-412

[11] Gozali Lina, Kristina Helena Juliana, et al. The improvement of block chain technology simulation in supply chain management (case study: Pesticide company). Scientific Reports, 2024, 14(1):3784

[12] Huang Ke, Mu Yi, Rezaeibagha Fatemeh, Zhang Xiaosong, Li Xiong, Cao Sheng. Monero with multi-grained redaction. IEEE Transactions on Dependable and Secure Computing, 2023, 21(1):241-253

[13] Fan Shulin, Hua Zhichao, Xia Yubin, Chen Haibo. Xputee: A high-performance and practical heterogeneous trusted execution environment for gpus. ACM Transactions on Computer Systems, 2025, 43(1-2):2

[14] Mo Fan, Tarkhani Zahra, Haddadi Hamed. Machine learning with confidential computing: A systematization of knowledge. ACM Computing Surveys, 2024, 56(11):1-40

[15] Yao Andrew C. Theory and application of trapdoor functions// Proceedings of the 23rd Annual Symposium on Foundations of Computer Science (SFCS 1982). Chicago, USA, 1982: 80-91

[16] He Fan, Xin Xiangjun, Li Chaoyang, Li Fagen. Security

analysis of the semi-quantum secret-sharing protocol of specific bits and its improvement. Quantum Information Processing, 2024, 23(2):51

[17] Ayeelyan John, Utomo Sapdo, Rouniyar Adarsh, Hsu Hsiu-Chun, Hsiung Pao-Ann. Federated learning design and functional models: Survey. Artificial Intelligence Review, 2025, 58(1):1-38

[18] Ji Xiaoyu, Li Junru, Song Yifan. Linear-communication asynchronous complete secret sharing with optimal resilience//Proceedings of the 44th Annual International Cryptology Conference (CRYPTO 2024). Santa Barbara, USA, 2024: 418-453

[19] Al Qahtani Elham, Story Peter, Shehab Mohamed. The impact of risk appeal approaches on users' sharing confidential in formation//Proceedings of the 2024 ACM Conference on Human Factors in Computing Systems. Hawaii, USA, 2024: 1-21

[20] Owusu Emmanuel, Guajardo Jorge, McCune Jonathan, Newsome Jim, Perrig Adrian, Vasudevan Amit. Oasis: On achieving a sanctuary for integrity and secrecy on untrusted platforms//Proceedings of the 2013 ACM SIGSAC Conference on Computer and Communications Security (CCS 2013). Berlin, Germany, 2013: 13-24

[21] Sadeghi Ahmad-Reza, Wachsmann Christian, Waidner Michael. Security and privacy challenges in industrial internet of things//Proceedings of the The 52nd Annual Design Automation Conference 2015. San Francisco, USA, 2015: 1-6

[22] Tramer Florian, Boneh Dan. Slalom: Fast, verifiable and private execution of neural networks in trusted hardware//Proceedings of the International Conference on Learning Representations. New Orleans, USA, 2019: Online only

[23] Ng LKL, SSM Chow, APY Woo, DPH Wong, Y Zhao. Goten: Gpu-outsourcing trusted execution of neural network training//Proceedings of the AAAI Conference on Artificial Intelligence. online 2021: 14876-14883

[24] Tran Anh-Tu, Luong The-Dung, Huynh Van-Nam. A comprehensive survey and taxonomy on privacy-preserving deep learning. Neurocomputing, 2024, 576:127345

[25] Mohassel Payman, Zhang Yupeng. Secureml: A system for scalable privacy-preserving machine learning//Proceedings of the 38th IEEE Symposium on Security & Privacy (SP 2017). JoseSan, USA, 2017: 19-38

[26] Gilad-Bachrach Ran, Dowlin Nathan, Laine Kim, Lauter Kristin, Naehrig Michael, Wernsing John. Cryptonets: Applying neural networks to encrypted data with high throu ghput and accuracy//Proceedings of the 33rd International Conference on Machine Learning (ICML 2016). New York, USA, 2016: 201-210

[27] Otoum Yazan, Gottimukkala Navya, Kumar Neeraj, Nayak Amiya. Machine learning in metaverse security: Current solutions and future challenges. ACM Computing Surveys, 2024, 56(8):1-36

[28] Juvekar Chiraag, Vaikuntanathan Vinod, Chandrakasan Anantha. {gazelle}: A low latency framework for secure neural network inference//Proceedings of the 27th USENIX security symposium (USENIX security 18). Baltimore, USA, 2018: 1651-1669

[29] Chen Jiasi, Ran Xukan. Deep learning with edge computing: A review. Proceedings of the IEEE, 2019, 107(8):1655-1674

[30] Nayan Tushar, Guo Qiming, Al Duniawi Mohammed, Botacin Marcus, Uluagac Selcuk, Sun Ruimin. {sok}: All you need to know about {on-device}{ml} model extraction-the gap between research and practice//Proceedings of the 33rd USENIX Security Symposium (USENIX Security'24). Philadelphia, USA, 2024: 5233-5250

[31] Udendhran R., Balamurugan M. Retracted article: Towards secure deep learning architecture for smart farming-based applications. Complex & Intelligent Systems, 2021, 7(2): 659-666

[32] Keller Marcel, Sun Ke. Secure quantized training for deep learning//Proceedings of the 39th International Conference on Machine Learning (ICML 2022). Baltimore, USA, 2022: 10912-10938

[33] Singh Maninderpal, Aujla Gagangeet Singh, Bali Rasmeet Singh. A deep learning-based blockchain mechanism for secure internet of drones environment. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(7):4404-4413

[34] Pang Qi, Yuan Yuanyuan, Wang Shuai. Mpcdiff: Testing and repairing mpc-hardened deep learning models//Proceedings of the 31st Annual Network and Distributed System Security Symposium. San Diego, USA, 2024: 716-734

[35] Asheralieva Alia, Niyato Dusit. Secure and efficient coded multi-access edge computing with generalized graph neural networks. IEEE Transactions on Mobile Computing, 2022, 22(9):5504-5524

[36] Mao Shuoyu, Yang Xinsong, Wang Maolin, Zheng Wei Xing. Secure stabilization of switched t-s fuzzy systems with mixed delay via mode-dependent event-triggered control. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2023, 54(1):255-264

[37] Garg Radhika, Yang Kang, Katz Jonathan, Wang Xiao. Scalable mixed-mode mpc//Proceedings of the 45th IEEE Symposium on Security and Privacy (IEEE S&P 2024). San Francisco, USA, 2024: 523-541

[38] Yuan Shougang, Awad Amro, Zhou Huiyang. Delta counter: Bandwidth-efficient encryption counter representation for secure gpu memory. IEEE Transactions on Dependable and Secure Computing, 2024, 22(1):101-113

[39] Shoup Victor, Smart Nigel P. Lightweight asynchronous verifiable secret sharing with optimal resilience. Journal of Cryptology, 2024, 37(3):27

[40] Puddu Ivan, Schneider Moritz, Lain Daniele, Boschetto Stefano, Srdjanapkun. On (the lack of) code confidentiality in trusted execution environments//Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP). San Francisco, USA, 2024: 4125-4142

[41] Leslie-Hurd Rebekah, Caspi Dror, Fernandez Matthew. Verifying linearizability of intel® software guard extensions//

Proceedings of the International Conference on Computer Aided Verification. Cham, 2015: 144-160

[42] Cheng Pau-Chen, Ozga Wojciech, et al. Intel tdx demystified: A top-down approach. ACM Computing Surveys, 2024, 56(9): 1-33

[43] Intel Intel. And ia-32 architectures software developer's manual. Volume 3B: system programming guide, Part, 2011, 2(11):0-40

[44] Baumann Andrew, Peinado Marcus, Hunt Galen. Shielding applications from an untrusted cloud with haven. ACM Transactions on Computer Systems (TOCS), 2015, 33(3):1-26

[45] Ohrimenko Olga, Schuster Felix, et al. Oblivious {multi-party} machine learning on trusted processors//Proceedings of the 25th USENIX Security Symposium (USENIX Security'16). Austin, USA, 2016: 619-636

[46] Hunt Tyler, Song Congzheng, Shokri Reza, Shmatikov Vitaly, Witchel Emmett. Chiron: Privacy-preserving machine learning as a service. arXiv preprint arXiv: 180305961, 2018, online only

[47] Park Saerom, Kim Seongmin, Yeon-supLim. Fairness audit of machine learning models with confidential computing// Proceedings of the The ACM Web Conference 2022. Online, 2022: 3488-3499

[48] Xiong Wenjie, Ke Liu, et al. Accelerating confidential recommendation model inference with near-memory processing. IEEE Transactions on Dependable and Secure Computing, 2025, 22(4):3580-3586

[49] Zhang Ziqi, Gong Chen, et al. No privacy left outside: On the (in-) security of tee-shielded dnn partition for on-device ml// Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP). San Francisco, USA, 2024: 3327-3345

[50] Sun Yu, Xiong Gaojian, Liu Jianhua, Liu Zheng, Cui Jian. Tsqp: Safeguarding real-time inference for quantization neural networks on edge devices//Proceedings of the 2025 IEEE Symposium on Security and Privacy (SP). San Francisco, USA, 2025: 2114-2132

[51] Han Zhaoxing, Hu Chengyu, Li Tongyaqi, Qi Qingqiang, Tang Peng, Guo Shanqing. Subgraph-level federated graph neural network for privacy-preserving recommendation with meta-learning. Neural Networks, 2024, 179:106574

[52] Xu Wei, Zhu Hui, et al. Tonn: An oblivious neural network prediction scheme with semi-honest tee. IEEE Transactions on Information Forensics and Security, 2024, 19:7335-7348

[53] Tsai Chia-Che, Porter Donald E., Vij Mona. {graphene-sgx}: A practical library {os} for unmodified applications on {sgx}// Proceedings of the 2017 USENIX annual technical conference (USENIX ATC 17). Santa Clara, USA., 2017: 645-658

[54] Will Newton C., Maziero Carlos A. Intel software guard extensions applications: A survey. ACM Computing Surveys, 2023, 55(14s):1-38

[55] Costan Victor, Lebedev Ilia, Devadas Srinivas. Sanctum: Minimal hardware extensions for strong software isolation// Proceedings of the 25th USENIX Security Symposium (USENIX Security 16). Austin, USA, 2016: 857-874

[56] Lee Dayeol, Kohlbrenner David, Shinde Shweta, Asanovi Krste, Song Dawn. Keystone: An open framework for architecting trusted execution environments//Proceedings of the 15th European Conference on Computer Systems (EuroSys '20. Heraklion, Greece, 2020: 1-16

[57] Kuvaiskii Dmitrii, Stavrakakis Dimitrios, Qin Kailun, Xing Cedric, Bhatotia Pramod, Vij Mona. Gramine-tdx: A lightweight os kernel for confidential vms//Proceedings of the 31st ACM SIGSAC Conference on Computer and Communications Security. Salt Lake City, USA, 2024: 4598-4612

[58] Fu Zhe, Sha Mo, et al. Encchain: Enhancing large language model applications with advanced privacy preservation techniques. Proceedings of the VLDB Endowment, 2024, 17(12):4413-4416

[59] Ivanov Andrei, Rothenberger Benjamin, Dethise Arnaud, Canini Marco, Hoefler Torsten, Perrig Adrian. {sage}: Software-based attestation for {gpu} execution//Proceedings of the USENIX Annual Technical Conference (USENIX ATC'23). BostonSheraton, USA, 2023: 485-499

[60] Oygenblik David, Yagemann Carter, Zhang Joseph, Mastali Arianna, Park Jeman, Saltaformaggio Brendan. {ai} psychiatry: Forensic investigation of deep learning networks in memory images//Proceedings of the 33rd USENIX Security Symposium (USENIX Security 24). Philadelphia, USA, 2024: 1687-1704

[61] Niu Yue, Ali Ramy E., Prakash Saurav, Avestimehr Salman. All rivers run to the sea: Private learning with asymmetric flows//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2024. Seattle, USA, 2024: 12353-12362

[62] Allaart Corinne, Amiri Saba, et al. Private and secure distributed deep learning: A survey. ACM Computing Surveys, 2024, 57(4):1-43

[63] Sun Haochen, Bai Tonghe, Li Jason, Zhang Hongyang. Zkdl: Efficient zero-knowledge proofs of deep learning training. IEEE Transactions on Information Forensics and Security, 2024, 20: 914-927

[64] Ghodsi Zahra, Gu Tianyu, Garg Siddharth. Safetynets: Verifiable execution of deep neural networks on an untrusted cloud//Proceedings of the The 31st Annual Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 4672-4681

[65] Quoc Do Le, Gregor Franz, Arnautov Sergei, Kunkel Roland, Bhatotia Pramod, Fetzer Christof. Securetf: A secure tensorflow framework//Proceedings of the 21st International Middleware Conference (Middleware'20). Delft, The Netherlands, 2020: 44-59

[66] Cheng Pau-Chen, Eykholt Kevin, et al. Deta: Minimizing data leaks in federated learning via decentralized and trustworthy aggregation//Proceedings of the Proceedings of the Nineteenth European Conference on Computer Systems. Athens, Greece, 2024: 219-235

[67] Jang Insu, Tang Adrian, Kim Taehoon, Sethumadhavan Simha, Huh Jaehyuk. Heterogeneous isolated execution for commodity gpus//Proceedings of the Architectural Support for

Programming Languages and Operating Systems. Providence RI, USA, 2019: 455-468

[68] Mai Haohui, Zhao Jiacheng, et al. Honeycomb: Secure and efficient {gpu} executions via static validation//Proceedings of the 17th USENIX Symposium on Operating Systems Design and Implementation (OSDI'23). Boston, USA, 2023: 155-172

[69] Li Mengyuan, Wilke Luca, Wichelmann Jan, Eisenbarth Thomas, Teodorescu Radu, Zhang Yinqian. A systematic look at ciphertext side channels on amd sev-snp//Proceedings of the 2022 IEEE Symposium on Security & Privacy (43rd). San Francisco, USA, 2022: 337-351

[70] Boyle Elette, Gilboa Niv, Ishai Yuval, Nof Ariel. Secure multiparty computation with sublinear preprocessing//Proceedings of the 41st Annual International Conference on the Theory and Applications of Cryptographic Techniques. Trondheim, Norway, 2022: 427-457

[71] Liu Ziyu, Zhou Tong, Luo Yukui, Xu Xiaolin. Tbnet: A neural architectural defense framework facilitating dnn model protection in trusted execution environments//Proceedings of the 61st ACM/IEEE Design Automation Conference (DAC'24). FranciscoSan, USA, 2024: 1-6

[72] Guan Le, Liu Peng, et al. Trustshadow: Secure execution of unmodified applications with arm trustzone//Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services (ACM MobiSys 2017). New York, USA, 2017: 488-501

[73] Van Nostrand Peter M., Kyriazis Ioannis, Cheng Michelle, Guo Tian, Walls Robert J. Confidential deep learning: Executing proprietary models on untrusted devices. arXiv preprint arXiv:190810730, 2019, Online only

[74] Mo Fan, Shamsabadi Ali Shahin, et al. Darknetz: Towards model privacy at the edge using trusted execution environments//Proceedings of the 18th Annual International Conference on Mobile Systems, Applications and Services (MobiSys 2020). Toronto, Ontario, Canada, 2020: 161-174

[75] Liu Bingyan, Lv Nuoyan, Guo Yuanchun, Li Yawen. Recent advances on federated learning: A systematic survey. Neurocomputing, 2024, 597(7):128019

[76] Xie Xueshuo, Wang Haoxu, et al. Memory-efficient and secure dnn inference on trustzone-enabled consumer iot devices//Proceedings of the 2024 IEEE/CVF International Conference on Computer Communications (IEEE INFOCOM 2024). Vancouver, Canada, 2024: 2009-2018

[77] Deng Yunjie, Wang Chenxu, et al. Strongbox: A gpu tee on arm endpoints//Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS'22). AngelesLos, USA, 2022: 769-783

[78] Sridhara Supraja, Bertschi Andrin, Schlter Benedict, Kuhne Mark, Aliberti Fabio, Shinde Shweta. {acai}: Protecting accelerator execution with arm confidential computing architecture//Proceedings of the 33rd USENIX Security Symposium(USENIX Security'24). Philadelphia, USA, 2024: 3423-3440

[79] Zhu Jianping, Hou Rui, et al. Enabling rack-scale confidential computing using heterogeneous trusted execution environment//Proceedings of the 2020 IEEE Symposium on Security and Privacy (IEEE S&P 2020). San Francisco, USA, 2020: 1450-1465

[80] Wang Chenxu, Deng Yunjie, et al. Building a lightweight trusted execution environment for arm gpus. IEEE Transactions on Dependable and Secure Computing, 2023, 21(4):3801-3816

[81] Dhar Aritra, Thorens Clement, Lazier Lara Magdalena, Cavigelli Lukas. Guardain: Protecting emerging generative ai workloads on heterogeneous npu//Proceedings of the 46th IEEE Symposium on Security and Privacy. San Francisco, USA, 2025: 4155-4172

[82] Luo Weile, Fan Ruibo, Li Zeyu, Du Dayou, Wang Qiang, Chu Xiaowen. Benchmarking and dissecting the nvidia hopper gpu architecture//Proceedings of the 38th IEEE International Parallel & Distributed Processing Symposium (IPDPS 2024). San Francisco, USA, 2024: 656-667

[83] Tan Yifan, Tan Cheng, Mi Zeyu, Chen Haibo. Pipellm: Fast and confidential large language model services with speculative pipelined encryption//Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2025). Rotterdam, The Netherlands, 2025: 843-857

[84] Na Seonjin, Kim Jungwoo, Lee Sunho, Huh Jaehyuk. Supporting secure multi-gpu computing with dynamic and batched metadata management//Proceedings of the 38th IEEE International Symposium on High-Performance Computer Architecture (HPCA 2024). Edinburgh, UK, 2024: 204-217

[85] Temucin Yiltan Hassan, Schonbein Whit, Levy Scott, Sojoodi Amirhossein, Grant Ryan E., AhmadAfsahi. Design and implementation of mpi-native gpu-initiated mpi partitioned communication//Proceedings of the SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis. Atlanta, USA, 2024: 436-447

[86] Jin Zheming. Sum reduction with openmp offload on nvidia grace-hopper system//Proceedings of the SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis. Atlanta, USA, 2024: 1006-1013

[87] Ali Sijjad, Wadho Shuaib Ahmed, Yichiet Aun, Gan Ming Lee, Lee Chen Kang. Advancing cloud security: Unveiling the protective potential of homomorphic secret sharing in secure cloud computing. Egyptian Informatics Journal, 2024, 27: 100519

[88] Li Lele, Han Zhaowei, Li Zhihui, Guan Feiting, Zhang Li. Authenticable dynamic quantum multi-secret sharing based on the chinese remainder theorem. Quantum Information Processing, 2024, 23(2):46

[89] Liu Yang, Kang Yan, et al. Vertical federated learning: Concepts, advances, and challenges. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(7):3615-3634

[90] Damg\aard Ivan, Pastro Valerio, Smart Nigel, Zakarias Sarah. Multiparty computation from somewhat homomorphic

encryption//Proceedings of the 32nd Annual Cryptology Conference(CRYPTO 2012/Advances in Cryptology-CRYPTO 2012). Santa Barbara, USA, 2012: 643-662

[91] Keller Marcel, Orsini Emmanuela, Scholl Peter. Mascot: Faster malicious arithmetic secure computation with oblivious transfer//Proceedings of the 23rd ACM SIGSAC Conference on Computer and Communications Security (CCS' 16). Vienna, Austria, 016: 830-842

[92] Riazi Sadegh M, Samragh Mohammad, Chen Hao, Laine Kim, Lauter Kristin, Koushanfar Farinaz. Xonn: Xnor-based oblivious deep neural network inference//Proceedings of the 28th USENIX Security Symposium (USENIX Security' 19). ClaraSanta, USA, 2019: 1501-1519

[93] Nielsen Jesper Buus, Nordholt Peter Sebastian, Orlandi Claudio, Burra Sai Sheshank. A new approach to practical active-secure two-party computation//Proceedings of the 32nd Annual International Cryptology Conference (CRYPTO 2012). Santa Barbara, USA, 2012: 681-700

[94] Larraia Enrique, Orsini Emmanuela, Smart Nigel P. Dishonest majority multi-party computation for binary circuits//Proceedings of the 34th Annual International Cryptology Conference (CRYPTO 2014). Santa Barbara, USA, 2014: 495-512

[95] Knott Brian, Shobha Venkataraman, Awni Hannun, Shubho Sengupta, Mark Ibrahim, Laurens Maaten. Crypten: Secure multi-party computation meets machine learning//Proceedings of the 34th Conference of Advances in Neural Information Processing Systems,Online, 2021: 4961-4973

[96] Kaissis Georgios, Alexander Ziller, et al. End-to-end privacy preserving deep learning on multi-institutional medical imaging. Nature Machine Intelligence, 2021, 3(6):473-484

[97] Rathee Mayank, Shen Conghao, Wagh Sameer, Popa Raluca Ada. Elsa: Secure aggregation for federated learning with malicious actors//Proceedings of the 2023 IEEE Symposium on Security and Privacy (IEEE S&P 2023). San Francisco, USA, 2023: 1961-1979

[98] Rotaru Dragos, Smart Nigel P, Tanguy Titouan, Vercauteren Frederik, Wood Tim. Actively secure setup for SPDZ. Journal of Cryptology, 2022, 35(1):5

[99] Zhang Yansong, Chen Xiaojun, et al. Md-sonic: Maliciously-secure outsourcing neural network inference with reduced online communication. IEEE Transactions on Information Forensics and Security, 2025, 20:3534-3549

[100] Keller Marcel. Mp-spdz: A versatile framework for multi-party computation//Proceedings of the 27th ACM SIGSAC Conference on Computer and Communications Security (CCS 2020). Online, 2020: 1575-1590

[101] Damgaard Ivan, Keller Marcel, Larraia Enrique, Pastro Valerio, Scholl Peter, Smart Nigel P. Practical covertly secure mpc for dishonest majority--or: Breaking the SPDZ limits//Proceedings of the 18th European Symposium on Research in Computer Security (ESORICS 2013). Egham, UK, 2013: 1-18

[102] Araki Toshinori, Furukawa Jun, Lindell Yehuda, Nof Ariel, Ohara Kazuma. High-throughput semi-honest secure three-party computation with an honest majority//Proceedings of the 23rd ACM SIGSAC Conference on Computer and Communications Security (CCS 2016). Vienna, Austria, 2016: 805-817

[103] Keller Marcel, Scholl Peter, Smart Nigel P. An architecture for practical actively secure mpc with dishonest majority//Proceedings of the 20th ACM SIGSAC Conference on Computer and Communications Security (CCS 2013). Berlin, Germany, 2013: 549-560

[104] Bi Renwan, Xiong Jinbo, et al. Communication-efficient privacy-preserving neural network inference via arithmetic secret sharing. IEEE Transactions on Information Forensics and Security, 2024, 19:6722-6737

[105] Feng Qi, He Debiao, Liu Zhe, Wang Huaqun, Choo Kim-Kwang Raymond. Securenlp: A system for multi-party privacy-preserving natural language processing. IEEE Transactions on Information Forensics and Security, 2020, 15:3709-3721

[106] Zeng Wenxuan, Li Meng, Yang Haichuan, Lu Wen-jie, Wang Runsheng, Huang Ru. Copriv: Network/protocol co-optimization for communication-efficient private inference. Advances in Neural Information Processing Systems, 2023, 36: 78906-78925

[107] Browning Jacob, LeCun Yann. Language, common sense, and the winograd schema challenge. Artificial Intelligence, 2023, 325:104031

[108] Zeng Chenkai, He Debiao, Feng Qi, Yang Xiaolin, Luo Qingcai. Securegpt: A framework for multi-party privacy-preserving transformer inference in gpt. IEEE Transactions on Information Forensics and Security, 2024, 19:9480-9493

[109] Zhang Yuke, Chen Dake, Kundu Souvik, Liu Haomei, Peng Ruiheng, Beerel Peter A. C 2 pi: An efficient crypto-clear two-party neural network private inference//Proceedings of the 60th ACM/IEEE Design Automation Conference (DAC 2023). San Francisco, USA, 2026: 1-6

[110] Shen Liyan, Dong Ye, et al. Abnn2: Secure two-party arbitrary-bitwidth quantized neural network pr edictions//Proceedings of the 59th ACM/IEEE Design Automation Conference (DAC 2022). San Francisco, USA, 2022: 361-366

[111] Agrawal Nitin, Shahin Shamsabadi Ali, Kusner Matt J., Gascon Adria. Quotient: Two-party secure neural network training and prediction//Proceedings of the 26th ACM SIGSAC Conference on Computer and Communications Security (CCS 2019). London, UK, 2019: 1231-1247

[112] Wagh Sameer, Tople Shruti, Benhamouda Fabrice, Kushilevitz Eyal, Mittal Prateek, Rabin Tal. Falcon: Honest-majority maliciously secure framework for private deep learning. arXiv preprint arXiv:200402229, 2020, Online only

[113] Wang Bobo, Yang Hongwei, Hao Meng, Zhang Jiannan, He Hui, Zhang Weizhe. Seppdl: A secure and efficient privacy-preserving deep learning inference framework for autonomous driving. ACM Transactions on Autonomous and Adaptive Systems, 2025, 20(1):5

[114] Ma Junming, Zheng Yancheng, et al. {secretflow-spu}: A performant and {user-friendly} framework for {privacy-

preserving} machine learning//Proceedings of the 2023 USENIX Annual Technical Conference (USENIX ATC'23). Boston, USA, 2023: 17-33

[115] Demmler Daniel, Schneider Thomas, Zohner Michael. Aby-a framework for efficient mixed-protocol secure two-party computation//Proceedings of the Network and Distributed System Security Symposium (NDSS 2015). DiegoSan, USA, 2015: Online only

[116] Patra Arpita, Schneider Thomas, Suresh Ajith, Yalame Hossein. {aby2. 0}: Improved {mixed-protocol} secure {two-party} computation//Proceedings of the 30th USENIX Security Symposium (USENIX Security'21). Online, 2021: 2165-2182

[117] Mohassel Payman, Rindal Peter. Aby3: A mixed protocol framework for machine learning//Proceedings of the 25th ACM SIGSAC Conference on Computer and Communications Security (CCS 2018). Toronto, Canada, 2018: 35-52

[118] Mishra Pratyush, Lehmkuhl Ryan, Srinivasan Akshayaram, Zheng Wenting, Popa Raluca Ada. Delphi: A cryptographic inference system for neural networks//Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice (PPMLP'20). Online, 2020: 27-30

[119] Huang Zhicong, Lu Wen-jie, Hong Cheng, Ding Jiansheng. Cheetah: Lean and fast secure {two-party} deep neural network inference//Proceedings of the 31st USENIX Security Symposium (USENIX Security'22). Boston, USA, 2022: 809-826

[120] Rizomiliotis Panagiotis, Diou Christos, Triakosia Aikaterini, Kyrannas Ilias, Tserpes Konstantinos. Partially oblivious neural network inference. arXiv preprint arXiv: 221015189, 2022, Online only

[121] Samragh Mohammad, Hussain Siam, Zhang Xinqiao, Huang Ke, Koushanfar Farinaz. On the application of binary neural networks in oblivious inference//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021). Online, 2021: 4630-4639

[122] Chen Hanxiao, Li Hongwei, et al. Secbnn: Efficient secure inference on binary neural networks. IEEE Transactions on Information Forensics and Security, 2024, 19:10273-10286

[123] Pang Qi, Zhu Jinhao, Mollering Helen, Zheng Wenting, ThomasSchneider. Bolt: Privacy-preserving, accurate and efficient inference for transformers//Proceedings of the 45th IEEE Symposium on Security and Privacy (IEEE S&P 2024). San Francisco, USA, 2024: 4753-4771

[124] Marcolla Chiara, Sucasas Victor, Manzano Marc, Bassoli Riccardo, Fitzek Frank H. P., Aaraj Najwa. Survey on fully homomorphic encryption, theory, and applications. Proceedings of the IEEE, 2022, 110(10):1572-1609

[125] Xie Qipeng, Jiang Siyang, et al. Efficiency optimization techniques in privacy-preserving federated learning with homomorphic encryption: A brief survey. IEEE Internet of Things Journal, 2024, 11(14):24569-24580

[126] Paillier Pascal. Public-key cryptosystems based on composite degree residuosity classes//Proceedings of the 32nd Annual International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT'99). Prague, Czech Republic, 1999: 223-238

[127] Gentry Craig. A fully homomorphic encryption scheme. Faculty of Science, Stanford, CA, USA, 2009

[128] Brakerski Zvika, Gentry Craig, Vaikuntanathan Vinod. (leveled) fully homomorphic encryption without bootstrapping. ACM Transactions on Computation Theory (TOCT), 2014, 6(3):1-36

[129] Brakerski Zvika, Vaikuntanathan Vinod. Efficient fully homomorphic encryption from (standard) lwe. SIAM Journal on computing, 2014, 43(2):831-871

[130] Fan Junfeng, Vercauteren Frederik. Somewhat practical fully homomorphic encryption. Cryptology ePrint Archive, 2012, Online only

[131] Cheon Jung Hee, Kim Andrey, Kim Miran, Song Yongsoo. Homomorphic encryption for arithmetic of approximate numbers//Proceedings of the 23rd International Conference on the Theory and Applications of Cryptology and Information Security (ASIACRYPT 2017). Kong Hong, China, 2017: 409-437

[132] Naresh Vankamamidi S., Ayyappa D. Ppdnn-crp: Ckks-fhe enabled privacy-preserving deep neural network processing for credit risk prediction. Computational Economics, 2024, 66(3): 2619-2643

[133] Lee Yongwoo, Cheon Seonyoung, Kim Dongkwan, Lee Dongyoon, Kim Hanjun. Performance-aware scale analysis with reserve for homomorphic encryption//Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2024). JollaLa, USA, 2024: 302-317

[134] Ju Jae Hyung, Park Jaiyoung, et al. Neujeans: Private neural network inference with joint optimization of convolution and fhe bootstrapping//Proceedings of the 31st ACM SIGSAC Conference on Computer and Communications Security (CCS'24). Salt Lake City, USA, 2024: 4361-4375

[135] Liu Yan, Lai Jianxin, et al. Resbm: Region-based scale and minimal-level bootstrapping management for fhe via min-cut//Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS'25). Rotterdam, The Netherlands, 2025: 924-939

[136] Li Long, Lai Jianxin, et al. Ant-ace: An fhe compiler framework for automating neural network inference//Proceedings of the 23rd ACM/IEEE International Symposium on Code Generation and Optimization (CGO 2025). Las Vegas, USA, 2025: 193-208

[137] Cheon Jung Hee, Kang Minsik, Kim Taeseong, Jung Junyoung, YongdongYeo. Batch inference on deep convolutional neural networks with fully homomorphic encryption using channel-by-channel convolutions. IEEE Transactions on Dependable and Secure Computing, 2024, 22(2):1674-1685

[138] Cho Wonhee, Hanrot Guillaume, Kim Taeseong, Park Minje, Stehle Damien. Fast and accurate homomorphic softmax evaluation//Proceedings of the 31st ACM SIGSAC Conference

on Computer and Communications Security (CCS'24). Salt Lake City, USA, 2024: 4391-4404

[139] Ao Wei, Boddeti Vishnu Naresh. {autofhe}: Automated adaption of {cnns} for efficient evaluation over {fhe}// Proceedings of the 33rd USENIX Security Symposium (USENIX Security'24). Philadelphia, USA, 2024: 2173-2190

[140] Lou Qian, Jiang Lei. Hemet: A homomorphic-encryption-friendly privacy-preserving mobile neural network architecture// Proceedings of the 38th International Conference on Machine Learning (ICML 2021). Online, 2021: 7102-7110

[141] Chen Hanxiao, Hao Meng, et al. Guardhfl: Privacy guardian for heterogeneous federated learning//Proceedings of the 40th International Conference on Machine Learning (ICML 2023). Honolulu, USA, 2023: 4566-4584

[142] Yuan Jiangjun, Liu Weinan, Shi Jiawen, Li Qingqing. Approximate homomorphic encryption based privacy-preserving machine learning: A survey. Artificial Intelligence Review, 2025, 58(3):82

[143] Boemer Fabian, Lao Yixing, Cammarota Rosario, Wierzynski Casimir. Ngraph-he: A graph compiler for deep learning on homomorphically encrypted data//Proceedings of the 16th ACM International Conference on Computing Frontiers (CF'19). Alghero, Italy, 2019: 3-13

[144] Lloret-Talavera Guillermo, Jorda Marc, et al. Enabling homomorphically encrypted inference for large DNN models. IEEE Transactions on Computers, 2021, 71(5):1145-1155

[145] Wood Alexander, Najarian Kayvan, Kahrobaei Delaram. Homomorphic encryption for machine learning in medicine and bioinformatics. ACM Computing Surveys (CSUR), 2020, 53(4): 1-35

[146] Zimerman Itamar, Baruch Moran, Drucker Nir, Ezov Gilad, Soceanu Omri, Wolf Lior. Converting transformers to polynomial form for secure inference over homomorphic encryption//Proceedings of the 41st International Conference on Machine Learning (ICML 2024). Vienna, Austria, 2024: Online only

[147] Dathathri Roshan, Saarikivi Olli, et al. Chet: An optimizing compiler for fully-homomorphic neural-network inferencing// Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI 2019). Phoenix, USA, 2019: 142-156

[148] Sarkar Sreetama, Kundu Souvik, Beerel Peter A. Rlnet: Robust linearized networks for efficient private inference// Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024). Washington, USA, 2024: 244-253

[149] Lin Yu, Zhang Tianling, Mao Yunlong, Zhong Sheng. Crossnet: A low-latency mlaas framework for privacy-preserving neural network inference on resource-limited devices. IEEE Transactions on Dependable and Secure Computing, 2024, 22(2):1265-1280

[150] Gouert Charles, Mouris Dimitris, Tsoutsos Nektarios Georgios. Helm: Navigating homomorphic encryption through gates and lookup tables. IEEE Transactions on Information Forensics and Security, 2025, 20:2822-2835

[151] Cheon Seonyoung, Lee Yongwoo, et al. {dacapo}: Automatic bootstrapping management for efficient fully homomorphic encryption//Proceedings of the 33rd USENIX Security Symposium (USENIX Security'24). Pennsylvania, USA, 2024: 6993-7010

[152] Ren Xuanle, Chen Zhaohui, et al. Cham: A customized homomorphic encryption accelerator for fast matrix-vector product//Proceedings of the 60th ACM/IEEE Design Automation Conference (DAC'23). FranciscoSan, USA, 2023: 1-6

[153] Ebel Austin, Garimella Karthik, Reagen Brandon. Orion: A fully homomorphic encryption framework for deep learning// Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2025). Rotterdam, The Netherlands, 2025: 734-749

[154] Liu Qingxiu, Huang Qun, et al. Pp-stream: Toward high-performance privacy-preserving neural network inference via distributed stream processing//Proceedings of the 40th IEEE International Conference on Data Engineering (ICDE 2024). Utrecht, The Netherlands, 2024: 1492-1505

[155] Nam Kevin, Jung Heonhui, Oh Hyunyoung, Paek Yunheung. Affinity-based optimizations for tfhe on processing-in-dram// Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS'25). Rotterdam, The Netherlands, 2025: 16-31

[156] Wan Junping, Li Danjie, Fang Junbing, Jiang Zoe L. Lpfhe: Low-complexity polynomial cnns for secure inference over fhe// Proceedings of the 29th European Symposium on Research in Computer Security (ESORICS 2024). Bydgoszcz, Poland, 2024: 403-423

[157] Pachon Ricardo, Trefethen Lloyd N. Barycentric-remez algorithms for best polynomial approximation in the chebfun system. BIT Numerical Mathematics, 2009, 49:721-741

[158] Checri Marina, Sirdey Renaud, Boudguiga Aymen, Bultel Jean-Paul. On the practical cpa d security of "exact" and threshold fhe schemes and libraries//Proceedings of the 43rd Annual International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT 2024). Zurich, Switzerland, 2024: 3-33

[159] Chen Jingxue, Yan Hang, Liu Zhiyuan, Zhang Min, Xiong Hu, Yu Shui. When federated learning meets privacy-preserving computation. ACM Computing Surveys, 2024, 56(12):1-36

[160] Nguyen Truc, Thai My T. Preserving privacy and security in federated learning. IEEE/ACM Transactions on Networking, 2023, 32(1):833-843

[161] Qammar Attia, Karim Ahmad, Ning Huansheng, Ding Jianguo. Securing federated learning with blockchain: A systematic literature review. Artificial Intelligence Review, 2023, 56(5): 3951-3985

[162] KerSIC Vid, Karakati SaSo, Turkanovic Muhamed. On-chain

zero-knowledge machine learning: An overview and comparison. Journal of King Saud University-Computer and Information Sciences, 2024, 36(9):102207

[163] Chen Binyi, BenediktBunz, DanBoneh, Zhang Zhenfei. Hyperplonk: Plonk with linear-time prover and high-degree custom gates//Proceedings of the 42nd Annual International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT 2023). Lyon, France, 2023: 499-530

[164] Bitansky Nir, Chiesa Alessandro, Ishai Yuval, Ostrovsky Rafail, Paneth Omer. Succinct non-interactive arguments via linear interactive proofs. Journal of Cryptology, 2022, 35(3):15

[165] Ben-Sasson Eli, Bentov Iddo, Horesh Yinon, Riabzev Michael. Scalable, transparent, and post-quantum secure computational integrity. Cryptology ePrint Archive, Online,2018

[166] Feng Boyuan, Qin Lianke, Zhang Zhenfei, Ding Yufei, Chu Shumo. Zen: An optimizing compiler for verifiable, zero-knowledge neural network inferences. Cryptology ePrint Archive, Online, 2021

[167] Sun Haochen, Li Jason, Zhang Hongyang. Zkllm: Zero knowledge proofs for large language models//Proceedings of the 31st ACM SIGSAC Conference on Computer and Communications Security (CCS'24). Salt Lake City, USA, 2024: 4405-4419

[168] Lee Seunghwa, Ko Hankyung, Kim Jihye, Oh Hyunok. Vcnn: Verifiable convolutional neural network based on zk-snarks. IEEE Transactions on Dependable and Secure Computing, 2024, 21(4):4254-4270

[169] Chen Bing-Jyue, Waiwitlikhit Suppakit, Stoica Ion, Kang Daniel. Zkml: An optimizing system for ml inference in zero-knowledge proofs//Proceedings of the 19th European Conference on Computer Systems. Athens, Greece, 2024: 560-574

[170] Feng Boyuan, Wang Zheng, Wang Yuke, Yang Shu, Ding Yufei. Zeno: A type-based optimization framework for zero knowledge neural network inference//Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2024). La Jolla, USA, 2024: 450-464

[171] Lu Tao, Chen Yuxun, Wang Zonghui, Wang Xiaohang, Chen Wenzhi, Zhang Jiaheng. Batchzk: A fully pipelined gpu-accelerated system for batch generation of zero-knowledge proofs//Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems. Rotterdam, The Netherlands, 2025: 100-115

[172] Liu Tianyi, Xie Xiang, Zhang Yupeng. Zkcnn: Zero knowledge proofs for convolutional neural network predictions and accuracy//Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. Republic of Korea, 2021: 2968-2985

[173] Abbaszadeh Kasra, Pappas Christodoulos, Katz Jonathan, Papadopoulos Dimitrios. Zero-knowledge proofs of training for deep neural networks//Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security. Salt Lake City, USA, 2024: 4316-4330

[174] Fan Yongkai, Ma Kaile, Zhang Linlin, Liu Jiqiang, Xiong Naixue, Yu Shui. Vericnn: Integrity verification of large-scale cnn training process based on zk-snark. Expert Systems with Applications, 2024, 255:124531

[175] Rosenberg Michael, Mopuri Tushar, Hafezi Hossein, Miers Ian, Mishra Pratyush. Hekaton: Horizontally-scalable zksnarks via proof aggregation//Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS 2024). Salt Lake City, USA, 2024: 929-940

[176] Liu Lei, Feng Jie, et al. Blockchain-enabled secure data sharing scheme in mobile-edge computing: An asynchronous advantage actor-critic learning approach. IEEE Internet of Things Journal, 2020, 8(4):2342-2353

[177] Yu Yu, Xie Xiang. Privacy-preserving computation in the post-quantum era. National Science Review, 2021, 8(9):nwab115

[178] Zhang Rui, Liu Jian, Ding Yuan, Wang Zhibo, Wu Qingbiao, Ren Kui. "Adversarial examples" for proof-of-learning//Proceedings of the 43rd IEEE Symposium on Security and Privacy. San Francisco, USA, 2022: 1408-1422

[179] Pappas Christodoulos, Papadopoulos Dimitrios. Sparrow: Space-efficient zksnark for data-parallel circuits and applications to zero-knowledge decision trees//Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security. Salt Lake City, USA, 2024: 3110-3124

[180] Ghodsi Zahra, Javaheripi Mojan, Sheybani Nojan, Zhang Xinqiao, Huang Ke, Koushanfar Farinaz. Zprobe: Zero peek robustness checks for federated learning//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 4860-4870

[181] Ma Renwen, Hwang Kai, Li Mo, Miao Yiming. Trusted model aggregation with zero-knowledge proofs in federated learning. IEEE Transactions on Parallel and Distributed Systems, 2024, 35(11):2284-2296

[182] Li Tan, Cheng Samuel, Chan Tak Lam, Hu Haibo. A polynomial proxy model approach to verifiable decentralized federated learning. Scientific Reports, 2024, 14(1):28786

[183] Duan Haohua, Peng Zedong, Xiang Liyao, Hu Yuncong, Li Bo. A verifiable and privacy-preserving federated learning training framework. IEEE Transactions on Dependable and Secure Computing, 2024, 21(5):5046-5058

[184] Zhang Bingxue, Lu Guangguang, Wu Yuncheng, Ren Kunpeng, Zhu Feida. Zkfhed: A verifiable and scalable blockchain-enhanced federated learning system. IEEE Transactions on Knowledge and Data Engineering, 2025, 37(6): 3841-3854

[185] Ma Juan, Liu Hao, Zhang Mingyue, Liu Zhiming. Vpfl: Enabling verifiability and privacy in federated learning with zero-knowledge proofs. Knowledge-Based Systems, 2024, 299: 112115

[186] Wang Haodi, Jiang Tangyu, Guo Yu, Guo Fangda, Bie Rongfang, Jia Xiaohua. Label noise correction for federated learning: A secure, efficient and reliable realization//Proceedings of the 40th IEEE International Conference on Data

Engineering (ICDE 2024). Utrecht, The Netherlands, 2024: 3600-3612

[187] Samardzic Nikola, Langowski Simon, Devadas Srinivas, Sanchez Daniel. Accelerating zero-knowledge proofs through hardware-algorithm co-design//Proceedings of the 57th IEEE/ACM International Symposium on Microarchitecture (MICRO 2024. Austin, USA, 2024: 366-379

[188] Yang Yibin, Heath David, Hazay Carmit, Kolesnikov Vladimir, Venkitasubramaniam Muthuramakrishnan. Tight zk cpu: Batched zk branching with cost proportional to evaluated instruction//Proceedings of the 31st ACM SIGSAC Conference on Computer and Communications Security (CCS 2024). Salt Lake City, USA, 2024: 3095-3109

[189] Liu Tianyi, Zhang Zhenfei, Zhang Yuncong, Hu Wenqing, Zhang Ye. Ceno: Non-uniform, segment and parallel zero-knowledge virtual machine. Journal of Cryptology, 2025, 38(2):17

[190] Smahi Abla, Li Hui, et al. Bv-icvs: A privacy-preserving and verifiable federated learning framework for v2x environments using blockchain and zksnarks. Journal of King Saud University-Computer and Information Sciences, 2023, 35(6):101542

[191] Yin Xuefei, Zhu Yanming, Hu Jiankun. A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions. ACM Computing Surveys (CSUR), 2021), 54(6):1-36

[192] Yazdinejad Abbas, Dehghantanha Ali, Karimipour Hadis, Srivastava Gautam, Parizi Reza M. A robust privacy-preserving federated learning model against model poisoning attacks. IEEE Transactions on Information Forensics and Security, 2024, 19: 6693-6708

[193] Aouedi Ons, Vu Thai-Hoc, et al. A survey on intelligent internet of things: Applications, security, privacy, and future directions. IEEE Communications Surveys & Tutorials, 2024, 27(2):1238-1292

[194] Zhang Yancheng, Zheng Mengxin, Shang Yuzhang, Chen Xun, Lou Qian. Heprune: Fast private training of deep neural networks with encrypted data pruning//Proceedings of the 37th Annual Conference on Neural Information Processing Systems (NeurIPS 2024). Vancouver Convention Centre, Vancouver, Canada, 2024: 51063-51084

[195] Bartusek James, Raizes Justin. Secret sharing with certified deletion//Proceedings of the 44th Annual International Cryptology Conference (CRYPTO 2024). Santa Barbara, USA, 2024: 184-214

[196] Johnson Jeremy. New directions in cryptography. IEEE Transactions on Information Theory, 1976, 22(6):644-654

[197] Yao Andrew C. Protocols for secure computations//Proceedings of the 23rd Annual Symposium on Foundations of Computer Science (SFCS 1982). Chicago, Illinois, USA, 1982: 160-164

[198] Rathee Deevashwer, Rathee Mayank, et al. Cryptflow2: Practical 2-party secure inference//Proceedings of the 27th ACM Conference on Computer and Communications Security (CCS 2020). Florida, USA, 2020: 325-342

[199] Feng Jun, Wu Yefan, Sun Hong, Zhang Shunli, Liu Debin. Panther: Practical secure 2-party neural network inference. IEEE Transactions on Information Forensics and Security, 2025, 20:1149-1162

[200] Jawalkar Neha, Gupta Kanav, Basu Arkaprava, Chandran Nishanth, Gupta Divya, Sharma Rahul. Orca: Fss-based secure training and inference with gpus//Proceedings of the 45th IEEE Symposium on Security and Privacy (SP 2024). San Francisco, USA, 2024: 597-616

[201] Ye Mang, Shen Wei, Du Bo, Snezhko Eduard, Kovalev Vassili, Yuen Pong C. Vertical federated learning for effectiveness, security, applicability: A survey. ACM Computing Surveys, 2024, 57(9):Online only

[202] Li Jinguo, Yan Yan, Zhang Kai, Li Chunlin, Yuan Peichun. Fpcnn: A fast privacy-preserving outsourced convolutional neural network with low-bandwidth. Knowledge-Based Systems, 2024, 283:111181

[203] Fu Yu, Tong Yu, et al. Swift: Fast secure neural network inference with fully homomorphic encryption. IEEE Transactions on Information Forensics and Security, 2025, 20:2793-2806

[204] Bruggemann Andreas, Schick Oliver, Schneider Thomas, Suresh Ajith, Yalame Hossein. Don't eject the impostor: Fast three-party computation with a known cheater//Proceedings of the 45th IEEE Symposium on Security and Privacy (SP 2024). San Francisco, USA, 2024: 503-522

[205] Keller Marcel, Pastro Valerio, Rotaru Dragos. Overdrive: Making SPDZ great again//Proceedings of the 37th Annual International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT 2018). Tel Aviv, Israel, 2018: 158-189

[206] Lehmkuhl Ryan, Mishra Pratyush, Srinivasan Akshayaram, Popa Raluca Ada. Muse: Secure inference resilient to malicious clients//Proceedings of the 30th USENIX Security Symposium (USENIX Security'21). Vancouver, Canada, 2021: 2201-2218

[207] Chandran Nishanth, Gupta Divya, Obbattu Sai Lakshmi Bhavana, Shah Akash. {simc}: {ml} inference secure against malicious clients at {semi-honest } cost//Proceedings of the 31st USENIX Security Symposium (USENIX Security'22). Boston, USA, 2022: 1361-1378

[208] Wei Chengkun, Yu Ruijing, Fan Yuan, Chen Wenzhi, Wang Tianhao. Securely sampling discrete Gaussian noise for multi-party differential privacy//Proceedings of the 30th ACM SIGSAC Conference on Computer and Communications Security (CCS 2023). Copenhagen, Denmark, 2023: 2262-2276

[209] Fu Yucheng, Wang Tianhao. Benchmarking secure sampling protocols for differential privacy//Proceedings of the 31st ACM SIGSAC Conference on Computer and Communications Security (CCS 2024). Salt Lake City, USA, 2024: 318-332

[210] Wang Baocang, Chen Yange, et al. Privacy-preserving convolutional neural network classification scheme with multiple keys. IEEE Transactions on Services Computing, 2024, 17(1): 322-335

[211] Li Fabing, Zhai Yuanhao, Cai Shuangyu, Gao Mingyu. Seesaw: Compensating for nonlinear reduction with linear computations for private inference//Proceedings of the 41st

International Conference on Machine Learning (ICML 2024). Vienna, Austria, 2024: 29266-29277

[212] Krastev Aleksandar, Samardzic Nikola, Langowski Simon, Devadas Srinivas, Sanchez Daniel. A tensor compiler with automatic data packing for simple and efficient fully homomorphic encryption. Proceedings of the ACM on Programming Languages, 2024, 8(PLDI):126-150

[213] Samardzic Nikola, Sanchez Daniel. Bitpacker: Enabling high arithmetic efficiency in fully homomorphic encryption accelerators//Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2024). California, USA, 2024: 137-150

[214] He Jiaxing, Yang Kang, et al. Rhombus: Fast homomorphic matrix-vector multiplication for secure two-party inference//Proceedings of the 31st ACM SIGSAC Conference on Computer and Communications Security. Salt Lake City, USA, 2024: 2490-2504

[215] Xu Tianshi, Wu Lemeng, Wang Runsheng, Li Meng. Privcirnet: Efficient private inference via block circulant transformation. Advances in Neural Information Processing Systems, 2024, 37:111802-111831

[216] Shafiq Muhammad, Ren Lijing, Srivastava Gautam, Zhang Denghui, Bourouis Sami, Gadekallu Thippa Reddy. Building privacy-preserving medical text models with a pre-trained transformer. IEEE Internet of Things Journal, 2024, 12(9): 11529-11538

[217] Feng Qihua, Li Peiya, et al. Evit: Privacy-preserving image retrieval via encrypted vision transformer in cloud computing. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(8):7467-7483

[218] Tian Yonglin, Wang Jiangong, Wang Yutong, Zhao Chen, Yao Fei, Wang Xiao. Federated vehicular transformers and their federations: Privacy-preserving computing and cooperation for autonomous driving. IEEE Transactions on Intelligent Vehicles, 2022, 7(3):456-465

[219] Hao Meng, Li Hongwei, Chen Hanxiao, Xing Pengzhi, Xu Guowen, Zhang Tianwei. Iron: Private inference on transformers//Proceedings of the 36th Annual Conference on Neural Information Processing Systems (NeurIPS 2022). New Orleans, USA, 2022: 15718-15731

[220] Liang Zi, Wang Pinghui, et al. Merge: Fast private text generation//Proceedings of the 38th Annual AAAI Conference on Artificial Intelligence (AAAI-24). British Columbia, Canada, 2024: 19884-19892

[221] Lu Wen-jie, Huang Zhicong, et al. Bumblebee: Secure two-party inference framework for large transformers//Proceedings of the 45th IEEE Symposium on Security and Privacy (IEEE S&P 2024). San Francisco, USA, 2023: Online only

[222] Xu Tianshi, Lu Wen-jie, et al. Breaking the layer barrier: Remodeling private transformer inference with hybrid {ckks} and {mpc}//Proceedings of the 34th USENIX Security Symposium (USENIX Security' 25). Washington, USA, 2025: 2653-2672

[223] Wang Jianfeng, Yang Huazhong, Deng Shuwen, Li Xueqing. Cimsat: Exploiting sat analysis to attack compute-in-memory architecture defenses//Proceedings of the 31st ACM SIGSAC Conference on Computer and Communications Security (CCS 2024). Salt Lake City, USA, 2024: 3436-3450

[224] Schaefer Thomas J. The complexity of satisfiability problems//Proceedings of the 10th Annual ACM Symposium on Theory of Computing (STOC 1978). San Diego, USA, 1978: 216-226

[225] Madusanka Tharindu, Pratt-Hartmann Ian, Batista-Navarro Riza Theresa. Natural language satisfiability: Exploring the problem distribution and evaluating transformer-based language models//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024). Bangkok, Thailand, 2024: 15278-15294

[226] Alyahya Tasniem Nasser, Menai Mohamed El Bachir, HassanMathkour. On the structure of the boolean satisfiability problem: A survey. ACM Computing Surveys (CSUR), 2022, 55(3):1-34

[227] Wang Ziyu, Wu Yuting, Park Yongmo, Lu Wei D. Safe, secure and trustworthy compute-in-memory accelerators. Nature Electronics, 2024, 7(12):1086-1097

[228] Yu Shimeng, Jiang Hongwu, Huang Shanshi, Peng Xiaochen, Lu Anni. Compute-in-memory chips for deep learning: Recent trends and prospects. IEEE Circuits and Systems Magazine, 2021, 21(3):31-56

[229] Lu Xingyu, Basaran Umit Yigit, Guler Basak. Scalable multi-round multi-party privacy-preserving neural network training. IEEE Transactions on Information Theory, 2024, 70(11):8204-8236

[230] He Zhenhao, Korolija Dario, et al. {accl＋}: An {fpga-based} collective engine for distributed applications//Proceedings of the 18th USENIX Symposium on Operating Systems Design and Implementation (OSDI' 24). Santa Clara, USA, 2024: 211-231

[231] Ji Zhuoran, Zhang Zhiyuan, Xu Jiming, Ju Lei. Accelerating multi-scalar multiplication for efficient zero knowledge proofs with multi-gpu systems//Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2024). California, USA, 2024: 57-70

[232] Pippenger Nicholas. On the evaluation of powers and related problems//Proceedings of the 17th Annual Symposium on Foundations of Computer Science (SFCS 1976). Houston, USA, 1976: 258-263

[233] Zhou Hao, Liu Changxu, Yang Lan, Shang Li, Yang Fan. Rezk: A highly reconfigurable accelerator for zero-knowledge proof. IEEE Transactions on Circuits and Systems I: Regular Papers, 2024, 72(2):802-815

[234] Zhang Ye, Wang Shuo, et al. Pipezk: Accelerating zero-knowledge proof with a pipelined architecture//Proceedings of the 48th ACM/IEEE Annual International Symposium on Computer Architecture (ISCA 2021). Online, 2021: 416-428

[235] Liu Changxu, Zhou Hao, Yang Lan, Xu Jiamin, Dai Patrick, Yang Fan. Gypsophila: A scalable and bandwidth-optimized multi-scalar multiplication architecture//Proceedings of the 61st

ACM/IEEE Annual International Design Automation Conference (DAC 2024). San Francisco, USA, 2024: 1-6
[236] Qiu Pengcheng, Wu Guiming, et al. Msmac: Accelerating multi-scalar multiplication for zero-knowledge proof// Proceedings of the 61st ACM/IEEE Design Automation Conference (DAC 2024). San Francisco, USA, 2024: 1-6
[237] Krishnan Sakkarai Samy Hari, Vidhya Krishnan. Distributed arithmetic-fir filter design using approximate karatsuba multiplier and vlcsa. Expert Systems with Applications, 2024, 249:123488
[238] Yang Yongkui, Lu Zhenyan, Zeng Jingwei, Liu Xingguo, Qian Xuehai, Yu Zhibin. Falic: An fpga-based multi-scalar multiplication accelerator for zero-knowledge proof. IEEE Transactions on Computers, 2024, 73(12):2791-2804
[239] Ryffel Theo, Trask Andrew, et al. A generic framework for privacy preserving deep learning. arXiv preprint arXiv: 181104017, 2018, Online only
[240] Demelius Lea, Kern Roman, Trugler Andreas. Recent advances of differential privacy in centralized deep learning: A systematic survey. ACM Computing Surveys, 2025, 57(6):1-28
[241] Liu Yang, Fan Tao, Chen Tianjian, Xu Qian, Yang Qiang. Fate: An industrial grade platform for collaborative learning with data protection. Journal of Machine Learning Research, 2021, 22(226):1-6
[242] Bonawitz Keith, Eichner Hubert, et al. Towards federated learning at scale: System design//Proceedings of the 1st Conference on Machine Learning and Systems (MLSys 2019). California, USA, 2019: 374-388
[243] He Chaoyang, Li Songze, et al. Fedml: A research library and benchmark for federated machine learning. arXiv preprint arXiv: 200713518, 2020, Online Only
[244] Beutel Daniel J., Topal Taner, et al. Flower: A friendly federated learning research framework. arXiv preprint arXiv: 200714390, 2020, Online only
[245] Caldas Sebastian, Duddu Sai Meher Karthik, et al. Leaf: A benchmark for federated settings. arXiv preprint arXiv: 181201097, 2018, Online only
[246] Xie Yuexiang, Wang Zhen, et al. Federatedscope: A flexible federated learning platform for heterogeneity. arXiv preprint arXiv:220405011, 2022, Online only
[247] Tang Yehui, Han Kai, Guo Jianyuan, Xu Chang, Xu Chao, Wang Yunhe. Ghostnetv2: Enhance cheap operation with long-range attention//Proceedings of the 36th Annual Conference on Neural Information Processing Systems (NeurIPS 2022). Louisiana, USA, 2022: 9969-9982
[248] Riedel Pascal, Schick Lukas, Schwerin Reinhold, Reichert Manfred, Schaudt Daniel, Hafner Alexander. Comparative analysis of open-source federated learning frameworks-a literature-based survey and review. International Journal of Machine Learning and Cybernetics, 2024, 15(11):5257-5278
[249] Li Qinbin, Wen Zeyi, et al. A survey on federated learning systems: Vision, hype and reality for data privacy and protection. IEEE Transactions on Knowledge and Data Engineering, 2021, 35(4):3347-3366
[250] Urmonov Odilbek, Sajid Shoaib, Aziz Zafar, Kim HyungWon. Federated object detection scenarios for intelligent vehicles: Review, case studies, experiments and discussions. IEEE Transactions on Intelligent Vehicles, 2024, early access:1-30
[251] Wang Xiangwen, Quinn Derek, Moody Thomas S., Huang Meilan. Aldele: All-purpose deep learning toolkits for predicting the biocatalytic activities of enzymes. Journal of Chemical Information and Modeling, 2024, 64(8):3123-3139
[252] Zhu Xinghua, Wang Jianzong, Hong Zhenhou, Xiao Jing. Empirical studies of institutional federated learning for natural language processing//Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020. Online, 2020: 625-634
[253] Singh Moirangthem Biken, Singh Himanshu, Pratap Ajay. Energy-efficient and privacy-preserving blockchain based federated learning for smart healthcare system. IEEE Transactions on Services Computing, 2023, 17(5):2392-2403
[254] Ye Rui, Ge Rui, et al. Fedllm-bench: Realistic benchmarks for federated learning of large language models//Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS 2024). British Columbia, Canada, 2024: 111106-111130
[255] Zeng Yan, Huang Chengchuang, et al. Efficientmoe: Optimizing mixture-of-experts model training with adaptive load balance. IEEE Transactions on Parallel and Distributed Systems, 2025, 34(4):677-688

WANG Bo-Bo, Ph. D. candidate. His main research interests include confidential computing and high-performance computing.

YANG Hong-Wei, Ph. D., associate researcher. His research interests include privacy computing and federated learning.

HAO Meng, Ph. D., associate professor. His research interests include high-performance computing.

HE Hui, Ph. D., professor, Ph. D. supervisor. Her research interests include network measurement and network security.

ZHANG Wei-Zhe, Ph. D., professor, Ph. D. supervisor. His research interests include information security and system architecture.

Background

This paper focuses on the study of neural network training and inference techniques based on confidential computing. The exponential growth of data-driven artificial intelligence has catalysed transformative developments across numerous domains. However, this progress has simultaneously raised critical concerns regarding privacy preservation and data security. The implications of these concerns are far-reaching: from the inadvertent exposure of personal interests and behavioural patterns to the potential leakage of corporate trade secrets and threats to national security. Once private data is misappropriated, the resulting damage can be both extensive and irreparable. Confidential computing technologies—through hardware-based trusted execution environments or cryptographic safeguards—ensure that data remains protected throughout its entire processing lifecycle, achieving the ideal of "usable yet invisible" data. As such, confidential computing offers a compelling pathway towards addressing these increasingly pressing challenges.

In response to these challenges, both academic and industrial communities have proposed a diverse range of confidential-computing-based solutions for neural network training and inference. Nevertheless, substantial obstacles persist in the path to practical deployment. Hardware-based schemes are often constrained by limitations in computational throughput, memory capacity, and compatibility with existing systems and software stacks. Meanwhile, software-based approaches typically involve complex secure protocols characterised by significant computational and communication overheads, which render large-scale adoption impractical under current infrastructural conditions.

This paper presents a comprehensive and in-depth survey of recent research on confidential-computing-based neural network training and inference. Through rigorous comparative analysis, we systematically synthesise pioneering and representative contributions across various subfields, while also offering critical insights into a series of influential studies that have emerged within the past two years. To provide a clear and structured understanding of the current research landscape, we adopt a classification framework supported by illustrative figures and summary tables. Building upon this foundation, we identify key challenges, outline emerging trends, and propose forward-looking research directions aimed at guiding the next phase of development in privacy-preserving AI. It is our intention that this work serve as a valuable reference for researchers and practitioners seeking to advance the state of the art in confidential and trustworthy machine learning systems.

This work was supported in part by the National Key R&D Program of China (Grant No. 2023YFB4503205), the National Natural Science Foundation of China (Grant No. U22A2036, 62202123, 62472122), the Natural Science Foundation of Heilongjiang Province (Grant No. LH2024F022), and the Fundamental Research Funds for the Central Universities (Grant No. HIT. NSFJG202433). The project is of strategic significance as it aims to develop confidential computing solutions for machine learning by leveraging existing hardware resources alongside cryptographic techniques and hardware-assisted security mechanisms, thereby ensuring the secure and effective use of sensitive data. Our team has a strong academic background in federated learning, confidential computing, and IoT security, and has achieved a number of influential results in these areas.