

基于图 Transformer 的门级硬件木马检测方法

徐宇航, 于洪, 郑锐

(信息工程大学, 河南 郑州 450001)

摘要: 图神经网络能够有效学习并检测门级网表中恶意植入的硬件木马 (HT), 但现有方法通常基于同构、同工艺电路进行训练, 导致模型泛化能力受限。为此, 提出了基于图 Transformer 的门级 HT 检测方法。首先, 将门级网表转换为通用图结构, 设计多维特征提取模块捕捉节点的本征属性与拓扑信息。其次, 采用三层堆叠图 Transformer 进行特征编码, 捕捉节点间全局依赖关系, 融合多层次特征以增强表达能力。最后, 引入带梯度反转层的双任务分类器, 抑制电路异构性带来的域偏移。在 Trust-Hub 公共库的实验表明: 所提方法在新思 90 nm 通用库 SAED 上实现了 95.2% 的召回率, 且相较于主流方法在 F1 分数上有所提高; 在基于 TRIT 的大规模数据集上实现了 87.9% 的平均召回率和 85.0% 的平均 F1 分数, 能够有效提升 HT 检测模型在跨工艺或异构网表中的鲁棒性。

关键词: HT 检测; 门级网表; 图神经网络; 域对抗训练; Transformer

中图分类号: TP309

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.AQ25240

Gate-level hardware trojan detection method based on graph Transformer

Xu Yuhang, Yu Hong, Zheng Rui

Information Engineering University, Zhengzhou 450001, China

Abstract: Graph neural networks can effectively learn and detect maliciously implanted hardware trojans (HT) in gate-level netlists. However, existing methods were usually trained on isomorphic and same-process circuits, which resulted in the generalization ability of the models being limited by the process and structural characteristics of the training set, making it difficult to cope with the graph structure distribution shift caused by the differences between the target circuit and the training set. To address this issue, a gate-level HT detection method based on graph transformer was proposed. Firstly, the gate-level netlist was converted into a general graph structure, and a multi-dimensional feature extraction module was designed to capture the intrinsic attributes and topological information of nodes. Secondly, a three-layer stacked graph Transformer was adopted for feature encoding to capture the global dependencies between nodes, and multi-level features were fused to enhance the expression ability. Finally, a dual-task classifier with a gradient reversal layer was introduced to suppress the domain shift caused by circuit heterogeneity. Experiments at the Trust-Hub public library showed that the proposed method achieved a recall rate of 95.2% on the synopsys 90 nm general library SAED, and its F1-score was improved compared with existing mainstream

收稿日期: 2025-11-01; 修回日期: 2026-01-24

通信作者: 于洪, yuhong_3210@163.com

基金项目: 国家重点研发计划资助项目 (No.2022YFB4500900)

Foundation Item: The National Key Research and Development Program of China (No.2022YFB4500900)

methods. On the large-scale dataset based on TRIT, it achieved an average recall rate of 87.9% and an average F1-score of 85.0%. This method effectively improves the robustness of HT detection models in cross-process or heterogeneous netlists.

Key words: hardware trojan detection, gate-level netlist, graph neural network, domain adversarial training, Transformer

0 引言

随着现代集成电路（IC）设计全球化进程不断加快，硬件生产商日益倾向于将设计与制造环节外包给第三方。然而，行业内缺乏针对第三方知识产权核（3P-IP）的安全规范标准，使其逐渐成为硬件木马（hardware trojan, HT）的主要“栖息地”^[1]。HT是一种隐蔽的恶意组件，可通过修改原有设计、插入额外电路或利用现有电路漏洞等方式实现其恶意功能，如窃取敏感信息、破坏系统正常运行或提供后门访问权限等^[2]。门级网表是集成电路设计流程中的重要中间表示，由基本逻辑单元（如与门、或门、触发器等）及其互连关系构成。HT在门级网表中的植入通常具有高度隐蔽性，其结构常伪装为正常逻辑单元，或嵌入在深层次路径中，难以通过仿真或形式化方法全面覆盖。此外，木马节点在整体电路中所占比例极低（通常不足1%），导致检测任务面临严重的类别不平衡问题。因此，如何从复杂的门级结构中准确识别出木马节点，是当前硬件安全领域的关键挑战之一。为此，国内外学者在HT安全领域开展了广泛研究，现有的HT检测方法根据技术原理可分为3类：形式化验证^[3-4]、特征分析^[5]和机器学习检测^[6]。

基于形式化验证的检测方法采用严格的数学方法对电路设计进行正确性验证，凭借其数学完备性可达到理论上零漏检。然而，随着电路规模的不断扩大，其状态空间呈现指数增长，形式化验证与逻辑验证分别受到电路复杂度和检测时耗限制。此外，人为定义的“木马行为”难以涵盖繁多的HT类型，复杂木马仍可能超出当前模型的逻辑表达能力。基于特征分析的检测方法将电路抽象为一系列定量或定性特征，并利用规则推理、统计分析或机器学习算法判定是否存在木马，该类方法在中小规模电路和可解释场景中能够实现高效、低成本的检测。然而，该类方法用

于检测高隐蔽性木马或异构电路设计时需要重新调参，同时，其检测效率仍受限于电路规模和检测时耗。

随着机器学习技术的发展，基于深度学习的硅前HT检测方法因其成本低、检测效率高且不需要黄金电路参考等优点得到广泛认可。然而，传统神经网络对HT的敏感度仍依赖于人为手动构建的特征工程。此外，其学习能力受到传统模型限制，在训练中难以有效建模节点间复杂的拓扑关系与长程依赖，并且无法充分捕捉不同维度特征之间的高阶关联。因此，当应用在不同工艺或异构电路时存在难以泛化的问题，检测能力受限。

针对上述问题，基于近年来采用多头注意力机制的Transformer^[7]在全局依赖建模和长序列处理方面的卓越表现，本文结合将Transformer拓展到图神经网络（GNN）领域的图Transformer模型^[8-9]，提出一种基于图注意力机制的门级HT检测方法，该方法实现了在无黄金电路参考的情况下精准检测不同网表电路中存在的木马。具体而言，本文方法采用图神经网络模拟电路行为，通过图结构建模、特征提取和图神经网络分类技术，实现高效的HT检测。此外，本文使用两个数据集来评估检测模型，包括基于新思90 nm通用库的SAED数据集及基于系统250 nm库的TRIT数据集^[10-11]。实验结果显示：本文所提模型在SAED数据集中实现了95.2%的平均召回率以及91.2%的平均F1分数，相比目前最先进的学习模型召回率提高了7.1%，F1分数提高了0.2%；在TRIT数据集中，本文所提模型分别在组合逻辑类型HT检测中获得86.5%的平均召回率和81.3%的平均F1分数，在时序逻辑类型HT检测工作中获得89.2%的平均召回率以及88.7%的平均F1分数。综上所述，本文的主要贡献如下。

1) 改进了基于GNN的HT检测模型。在传统GNN中融合了多头注意力机制，使模型聚焦于

多个特征子空间,有效捕捉不同维度特征之间的重要关联,提高HT检测模型的鲁棒性。

2) 构建了一套非启发式特征工程。根据网表拓扑结构、门类型和全局位置信息构建多维度特征向量,而不针对HT构建任何启发式特征,提升模型对异构电路设计的泛用性。

3) 提出一种基于域对抗训练的域不变分类器(domain-invariant classifier, DIC)。采用木马检测和域分类的双任务并行架构,在域分类器中引入梯度反转机制,抑制其对不同电路设计的敏感性,从而缓解训练与测试数据来自不同电路设计所产生的域偏移问题。

1 相关工作

随着IC规模的不断增大,传统仿真测试难以穷尽所有可能状态,研究人员开始寻求数学方法形式化地证明电路是否满足规范,该方法同样被引入硬件安全领域。Fyrbiak等^[12]首次将图相似度算法引入硬件安全领域,在门级网表逆向工程、HT检测和混淆评估中验证了有效性。此外,梅波等^[13]提出了基于广义随机着色Petri网的分布式协同木马攻击模型,通过量化分析系统安全性指标,为防御策略的制定提供了理论支撑,但其均依赖纯拓扑特征而忽略语义信息,对代码级变种木马的泛化能力有限。武玲娟等^[14]利用电路内部可满足性无关项设计了高隐蔽性木马,在正常工作条件下与原始电路功能完全等价,从而逃避常规功能一致性验证与检测。Piccolboni等^[15]基于控制流子图匹配实现寄存器传输级(register transfer level, RTL)木马检测,支持参数化扩展以识别木马变种,但需预定义模式库且对未知木马类型适应性弱,在简单触发条件下误报率高。Chen等^[16]提出了一种面向第三方数字IP核的多层次特征分析方法,将触发器级结构特征分析与组合逻辑级信号翻转率分析相结合,实现了对硬件木马的快速检测,然而其特征仍停留在人工设计层面,无法自动学习跨设计域的深层模式,导致对乘法等正常大扇入结构存在误报。上述工作均基于已有木马构建相关方法库或数据库,其检测范围仅限于相应库中的已知HT。文献^[17]聚焦于结构特征与统计学习结合的木马检测,显著提升了HT检测效率,但均未有效建模

特征间动态交互关系与跨域泛化能力,在面对伪装性强或分布偏移的木马时鲁棒性不足。

近年来,图神经网络技术在社交网络、生物信息等领域取得显著成效,其结构建模能力也被引入HT检测任务。常见的GNN架构包括图卷积网络(GCN)^[18]、图注意力网络(GAT)^[19]和图采样聚合算法(GraphSAGE)^[20]等。门级网表天然的图结构为GNN提供了丰富的拓扑与语义信息,促使研究者探索更具表达力的检测范式。Yasaei等^[21]首次提出将GNN模型应用于RTL HT检测领域的方法,文献^[22]提出将GAT运用于检测模型,实现了对RTL与门级网表中未知HT的高效检测,但未深入研究注意力机制,史江义等^[23]使用GraphSAGE学习门级网表中的高维图特征以及相应节点特征,并探索了不同聚合方式以及数据平衡方法下的模型检测能力,文献^[24]在特征工程中引入调和中心性和权重优化算法提高检测质量。文献^[25]结合图卷积网络与图注意力网络用于RTL中的HT检测,但模型受特征向量表征能力限制,召回率仅为88.4%,存在较多未检测木马节点。

现有GNN方法在HT检测中仍面临3类局限。其一,长距离依赖建模不足。GCN的感受野受网络层数严格限制,需通过多层堆叠才能覆盖跨逻辑层级的路径,易导致梯度消失;GAT的注意力机制仅局限于一阶邻居,若堆叠多层则会引发过平滑问题,而HT的触发路径可跨越3~5级逻辑深度(如触发器→中间逻辑→有效载荷),传统方法难以完整捕捉其隐蔽关联。其二,异构图适应性薄弱。GCN、GAT均默认图结构同构且节点特征分布一致,但在实际HT检测中,训练集与测试集常来自不同工艺(如90 nm与250 nm)或设计风格(组合逻辑与时序逻辑),导致特征分布偏移,传统GNN缺乏自适应调整机制,泛化能力受限。其三,全局结构感知缺失。HT的异常模式不仅依赖局部节点特征,还与全图拓扑关联性强(如孤立的高介数中心性节点、异常聚集的子图),而GCN基于局部邻域聚合,GraphSAGE依赖随机采样,均难以全面捕捉全图节点的动态关联。

相比之下,图Transformer作为新兴图表示学习架构,在理论层面更加适配HT检测需求。

①长路径建模能力：通过多头自注意力机制，单层即可实现全图节点的直接信息交互，不需要依赖多层堆叠，能高效覆盖HT跨层级触发路径，避免梯度消失与过平滑问题。②异构图适配性：注意力权重可根据节点特征与拓扑关系自适应调整，不需要假设图结构同构，能自动区分不同工艺、设计风格的电路结构差异，从理论上缓解域偏移问题。③全局依赖捕捉：节点间注意力计算不受边缘连接限制，可直接建模任意节点对的关联强度，能精准捕捉从触发条件到恶意载荷的完整隐蔽路径，以及全图范围内的异常拓扑模式，如孤立恶意节点与关键信号节点的潜在关联。

2 本文方法

2.1 整体框架

本文方法的整体框架如图1所示，分为3个主要模块：特征提取模块、训练融合模块和域不变分类器。总体流程如下：首先，通过语法解析将原始门级网表转换为有向图结构，提取图中节点的拓扑属性、门类型和全局位置特征，并进行归一化预处理；其次，将特征输入三层图Transformer架构进行节点嵌入学习，通过多头注意力机制捕获全局依赖关系，并结合焦点损失（FocalLoss）函数优化极端不平衡样本的训练效果；最后，通过域不变分类器引入梯度反

转层，对抗训练以抑制工艺差异引起的域偏移，最终输出节点是否为HT的分类结果。下面对3个模块分别进行介绍。

2.2 特征提取模块

如图1所示，特征提取模块是实现电路结构精细化表征的核心环节，其功能是将原始门级网表转化为图2所示的结构化特征向量。特征向量的构建流程：通过Verilog语法解析原始门级网表，将基本逻辑门与时序单元转化为图节点、信号连接关系转化为有向边，形成通用图结构；从图结构中提取拓扑、结构、类型3类特征；特征经标准化预处理后，融合形成图2中所示的多维特征向量，作为后续特征编码模块的输入。

传统GNN特征通常只包括节点度、节点类型和单一的图结构特征，所提取的特征维度较低，并且通常只关注局部特征，适用于简单的图任务。学习检测HT需要更加关注电路设计中的深层次或隐蔽路径。为了更有效地捕捉HT的异常结构和隐蔽模式，需要设计多维度、多语义和结构感知的特征工程。

如图1所示，本文构建了一套完整的特征提取框架。

首先，通过解析将原始门级网表转化为一种通用的图数据结构。具体而言，采用Verilog语

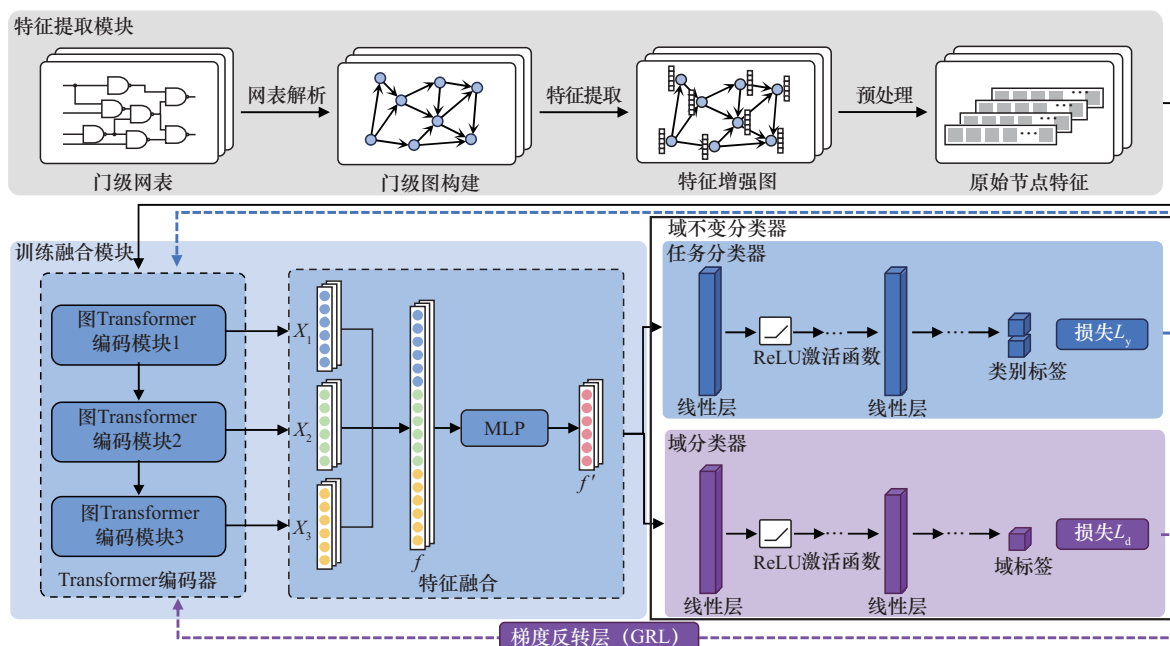


图1 本文方法的整体框架

法解析器提取门级网表中的结构信息，将基本逻辑门（AND/NAND等）、时序单元（D型触发器（DFF）/扫描D型触发器（SDFF）等）作为节点，并将模块实例（通用异步收发传输器（UART）等）排除，此外，通过信号追踪建立有向边，并且排除图中的环结构。

其次，通过多级处理流程实现电路结构的精细化表征。具体分为两点，一是在信号提取阶段，采用正则表达式解析 Verilog 文件中的 I/O 声明，将总线信号进行位级扩展，将多位信号等声明分解为离散信号，通过融合函数实现原始声明信号与扩展信号的智能合并，同时，利用集合运算消除冗余，保留信号顺序。二是在门级特征提取阶段，解析门实例化语句构建连接关系，提取拓扑、结构和类型3类关键特征，结构化特征向量表示如图2所示。其中，拓扑特征包括聚类系数（CC）、度中心性（DC）和介数中心性（BC），用于刻画节点在图中的连接紧密度与关键性，有助于识别异常聚集或中心性偏离的潜在木马结构；结构特征包括出度、入度、节点与主输入/输出的最小距离（PI/PO），能够反映节点在电路中的连接强度与位置重要性，由于木马常位于深层次或低活跃路径，此类特征可有效捕捉其结构异常；类型特征通过端口/门类型的独热编码区分功能单元，避免由门类型差异导致的误判。因此，通过多维特征融合，模型能够在后继特征编码模块增强对跨层级、跨工艺电路中木马特征的判别能力，提升检测的鲁棒性。

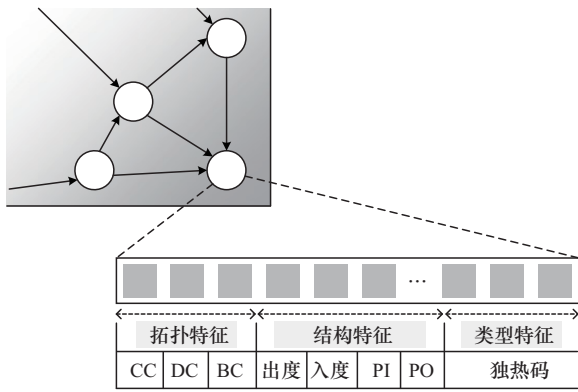


图2 结构化特征向量表示

最后，在预处理环节引入动态归一化机制，对节点入度、节点出度、PI、PO等数值特征进

行 Z-score 标准化，确保模型训练的稳定性，最终得到原始节点特征作为训练融合模块中图神经网络的输入。

2.3 训练融合模块

在完成网表图结构构建和节点特征提取后，将经过归一化处理的节点特征输入基于图 Transformer 的图神经网络模型进行节点嵌入学习和模型训练。由于硬件网表通常具有 3~5 级逻辑深度，因此本文采用 3 层 Transformer 架构进行特征提取，通过顺序执行 3 层图 Transformer 以捕捉和强化输入特征中的长距离依赖和全局信息，逐步提取和抽象特征，过滤冗余信息，从而生成更高层次、更抽象的特征表示。

传统注意力机制通常只能关注到序列中的局部特征，而本文采用的图 Transformer 是一种基于多头注意力机制的空域图卷积算法，多头注意力能够聚焦于不同的特征子空间，动态学习节点间的关系权重。该设计使模型能够有效捕捉电路图中的关键结构特征。图 Transformer 模块结构如图3所示，下面对其运算流程进行介绍。

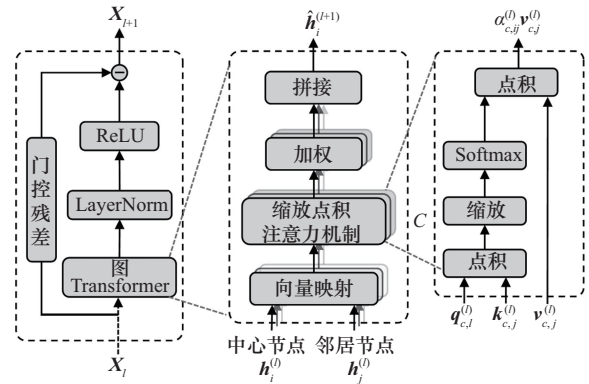


图3 图 Transformer 模块结构

1) 计算注意力分数：对于每一层 l ，对每个节点 i 及其邻居节点 $j \in N(i)$ ，通过 3 组权重矩阵计算第 c 个注意力头的查询向量 $q_{c,i}^{(l)}$ 、键向量 $k_{c,j}^{(l)}$ 和值向量 $v_{c,j}^{(l)}$ ，过程为：

$$q_{c,i}^{(l)} = W_{c,q}^{(l)} h_i^{(l)} + b_{c,q}^{(l)} \quad (1)$$

$$k_{c,j}^{(l)} = W_{c,k}^{(l)} h_j^{(l)} + b_{c,k}^{(l)} \quad (2)$$

$$v_{c,j}^{(l)} = W_{c,v}^{(l)} h_j^{(l)} + b_{c,v}^{(l)} \quad (3)$$

其中， $h_i^{(l)}$ 和 $h_j^{(l)}$ 分别表示第 l 层节点的输入特征表

示, W 和 b 为对应线性映射的可学习参数。通过计算 $q_{c,i}^{(l)}$ 和 $k_{c,j}^{(l)}$ 的点积并经过向量维度 d 缩放和Softmax归一化得到注意力权重 $\alpha_{c,ij}^{(l)}$, 如式(4)所示。

$$\alpha_{c,ij}^{(l)} = \text{softmax}\left(\frac{q_{c,i}^{(l)} \cdot k_{c,j}^{(l)}}{\sqrt{d}}\right) \quad (4)$$

2) 消息聚合: 对于每个注意力头, 计算邻居节点的值向量 $v_{c,j}^{(l)}$ 并加权求和, 得到节点 i 在第 $l+1$ 层的聚合结果 $\hat{h}_i^{(l+1)}$ 为

$$\hat{h}_i^{(l+1)} = \text{Concat}_{c=1}^C \left(\sum_{j \in N(i)} \alpha_{c,ij}^{(l)} v_{c,j}^{(l)} \right) \quad (5)$$

其中, C 为注意力头数, Concat表示多头输出的拼接操作, 可替换为使用平均操作, 如式(6)所示, 二者对模型检测能力的影响将在3.2节展开研究。

$$\hat{h}_i^{(l+1)} = \frac{1}{C} \sum_{c=1}^C \sum_{j \in N(i)} \alpha_{c,ij}^{(l)} v_{c,j}^{(l)} \quad (6)$$

3) 门控残差链接: 为防止深层网络中的过平滑, 引入门控机制融合当前层输出与残差项, 随后将聚合结果输入层归一化函数LayerNorm和激活函数ReLU, 得到最终输出 $h_i^{(l+1)}$, 具体运算过程为:

$$r_i^{(l)} = W_r^{(l)} h_i^{(l)} + b_r^{(l)} \quad (7)$$

$$\beta_i^{(l)} = \sigma(W_\beta^{(l)}[h_i^{(l)}; r_i^{(l)}]) \quad (8)$$

$$h_i^{(l+1)} = \text{ReLU}(\text{LayerNorm}((1 - \beta_i^{(l)})\hat{h}_i^{(l+1)} + \beta_i^{(l)}r_i^{(l)})) \quad (9)$$

其中, $r_i^{(l)}$ 表示由第 l 层输入特征得到的残差项, $W_r^{(l)}$ 、 $b_r^{(l)}$ 和 $W_\beta^{(l)}$ 为可学习参数, $\beta_i^{(l)}$ 对应门控系数, σ 为Sigmoid激活函数。将各层图Transformer模块的输出表示为3组特征向量矩阵, 分别为: $X_1\{h_1^1, h_2^1, h_3^1, \dots, h_n^1\}$ 、 $X_2\{h_1^2, h_2^2, h_3^2, \dots, h_n^2\}$ 和 $X_3\{h_1^3, h_2^3, h_3^3, \dots, h_n^3\}$ 。

将3层Transformer模块的特征输出至特征融合层进行融合和压缩, 生成更具表达力的统一特征表示, 用于后继分类任务。具体而言, 先将各图Transformer模块的输出在特征维度上拼接, 得到一个多尺度特征向量矩阵, 再通过多层感知机(multi-layer perceptron, MLP)对特征进行融合降维, 最终输出融合向量矩阵 f' 。具体为:

$$f = \text{Concat}(X_1, X_2, X_3) \quad (10)$$

$$f' = \text{MLP}(f) \quad (11)$$

综上所述, 基于图Transformer的特征编码

器通过引入多头注意力机制, 能够有效整合节点自身特征与全局拓扑结构信息, 实现对门级网表中长距离依赖关系的建模。其中, 多层堆叠结构不仅增强了模型对复杂电路结构的表达能力, 还通过逐层抽象提取出更具判别力的高层特征表示。

2.4 域不变分类器

在实际HT检测任务中, 训练数据(源域)与测试数据(目标域)通常来自不同的电路设计, 因此存在显著的分布差异, 即域偏移问题[26]。为了提升模型在跨设计、跨工艺情况下的检测能力, 一方面, 需要保证模型对木马的分类能力; 另一方面, 需要抑制模型对不同电路设计的敏感性。因此, 本文设计了一个基于域对抗训练的域不变分类器, 其核心思想是确保模型学习到的特征在不同电路设计(域)之间具有不变性, 从而提高模型在未知电路上的泛化能力。如图1所示, DIC采用双任务并行架构, 接收特征编码模块的输出, 分别输入到任务分类器和域分类器, 预测分类并计算损失值。

2.4.1 任务分类器

任务分类器 G_f 采用两层全连接网络, 负责木马节点检测, 经过Softmax归一化得到概率分布 $\hat{y}$, 过程如式(12)所示。

$$\hat{y} = G_f(f'; \theta_f) \quad (12)$$

其中, f' 为训练融合模块的输出, θ_f 表示任务分类器的权重矩阵参数。

由于木马分类任务中绝大多数电路节点为非木马节点, 极少数为木马节点。因此, 在任务分类中使用FocalLoss函数[27]代替常用的交叉熵损失函数来计算任务损失 L_f 。

$$L_f = \frac{1}{n_s} \sum_{i=1}^{n_s} \text{FocalLoss}(y_i, \hat{y}_i) \quad (13)$$

$$\text{FocalLoss}(p_t) = -\alpha(1 - p_t)^\gamma \ln(p_t) \quad (14)$$

其中, 参数 γ 作为调整因子控制模型对难分类样本的聚焦程度, 参数 α 则直接处理类别数量上的不平衡, n_s 为源域节点数, 置信度概率 p_t 由预测标签 $\hat{y}_i$ 和真实标签 y_i 确定。

$$p_t = \begin{cases} \hat{y}_i, & y_i = 1 \\ 1 - \hat{y}_i, & y_i = 0 \end{cases} \quad (15)$$

因此, 易区分样本的 p_t 较高, 在交叉熵损失前增加动态权重因子 $(1 - p_t)^\gamma$, 使其接近0, 自

动降低高置信度样本的损失贡献。同样，难区分样本由于 p_t 较小，模型可自动提高其损失贡献。

2.4.2 域分类器

域分类器 G_d 采用单层网络，通过梯度反转层（gradient reversal layer, GRL）^[28] 连接特征空间，域预测概率分布 $\hat{d}$ 为：

$$\hat{d} = G_d(f'; \theta_d) \quad (16)$$

其中， θ_d 表示域分类器的权重矩阵参数。

训练时，每个批量中抽取相同数量的源域样本和目标域样本，域分类任务中不存在类别不平衡问题，因此，本文采用二元交叉熵（balanced cross entropy, BCE）计算域损失 L_d 。

$$L_d = \frac{1}{n_s + n_t} \sum_{i=1}^{n_s + n_t} \text{BCE}(d_i, \hat{d}_i) \quad (17)$$

其中， n_s 、 n_t 分别为源域和目标域节点数量。

2.4.3 域对抗训练

由 L_f 和 L_d 可计算得到总损失 L 。

$$L = L_f + \beta L_d \quad (18)$$

其中，超参数 β 用于调节域损失占总损失的权重。

对于任务分类器权重矩阵参数 θ_f 、域分类器权重矩阵参数 θ_d 和特征提取器权重矩阵参数 θ_y ，遵循梯度下降参数更新规则，得到各组参数的更新公式。

$$\theta_y \leftarrow \theta_y - \eta \frac{dL}{d\theta_y} \quad (19)$$

$$\theta_d \leftarrow \theta_d - \eta \frac{dL}{d\theta_d} \quad (20)$$

$$\theta_f \leftarrow \theta_f - \eta \frac{dL}{d\theta_f} \quad (21)$$

其中， η 为学习率，决定整体更新步长的大小，根据式(18)，总损失对 θ_y 的梯度可以拆分为任务梯度和域梯度两部分。

$$\frac{dL}{d\theta_y} = \frac{dL_f}{d\theta_y} + \beta \frac{dL_d}{d\theta_y} \quad (22)$$

由于在计算 $\frac{dL_d}{d\theta_y}$ 的路径上插入了 GRL，该层可

在反向传播时反转域分类器梯度方向，驱使训练融合模块生成混淆域分类器的特征表示，从而减小源域与目标域的特征分布差异。因此，实际反向传播

结果为 $-\lambda \frac{dL_d}{d\theta_y}$ ，特征提取器的参数更新公式为：

$$\theta_y \leftarrow \theta_y - \eta \left(\frac{dL_f}{d\theta_y} - \beta \lambda \frac{dL_d}{d\theta_y} \right) \quad (23)$$

其中，超参数 λ 为梯度反转强度，控制特征提取器对域分类器的“欺骗”程度。

综上所述，DIC 通过对抗训练学习跨域可迁移特征，有效缓解了源域与目标域之间的分布差异问题。其中，通过任务分类器与域分类器的协同优化，模型能够在特征空间中学习到更具判别力且域不变的表示，从而在面对未知或异构网表时，依然保持较高的检测准确率与稳定性，提升模型在未知目标域上的泛化能力。

2.5 适配性分析

在 HT 检测任务中，图 Transformer 与域对抗训练结合的优势在于二者形成了功能互补，协同解决长距离依赖建模与域泛化困难。

一方面，HT 的隐蔽性通常在于其触发路径跨越多个逻辑层级，如时序木马的有限状态机（FSM）状态转移链、组合木马的多门级触发条件。传统 GNN 受限于局部邻域聚合机制，难以捕捉跨层级异常关联，而图 Transformer 的多头自注意力机制允许任意节点直接计算关联权重，能够一次性建模寄存器→组合逻辑→输出端口间的长距离依赖，定位触发点与载荷点之间的路径，且不需要依赖边连接或网络层数扩展，为后续分类提供强判别力的全局特征。

另一方面，跨工艺、跨设计场景下的电路异构性会导致特征分布偏移，单纯的图 Transformer 虽然能够自适应学习节点关联，但仍可能过度拟合源域的工艺特性特征，如特定工艺下的门类型分布、拓扑偏好等，导致在目标域上的泛化性能下降。域对抗训练通过梯度反转层迫使特征提取器学习域不变特征，即剥离工艺、设计风格等无关差异，保留木马节点的本质结构特征，如异常聚集性和长路径关联性。通过全局依赖建模与域不变特征提取的协同机制，模型能够精准识别木马的隐蔽结构模式，同时适应不同电路的异构性，符合跨工艺、跨设计的 HT 检测需求。

3 实验

本节将对本文实验所用数据集进行解释，首先，通过 SAED 网表库进行模型寻优，详细评估检

测模型性能。其次，与当前先进方法进行对比，具体包括基于图神经网络的HT检测方法（Netlist-GNN^[22]）、基于图采样聚合算法的门级HT检测方法（GraphSAGE^[23-24]）和基于对抗训练的鲁棒HT检测方法（R-HTDetector^[29]）。最后，基于组合触发型硬件木马测试集（TRIT-TC）和时序触发型硬件木马测试集（TRIT-TS）进行鲁棒性测试，进一步验证本文方法的有效性。

3.1 实验设置

3.1.1 环境设置

本文实验模型由PyTorch及其图神经网络库（PyTorch geometric, PyG）搭建，其中PyG提供了丰富的图神经网络层（如GCNConv、GATConv、TransformerConv等），以及高效的稀疏张量操作，能够用于快速构建和训练图神经网络模型。实验运行的硬件平台为NVIDIA GeForce RTX4070tis, AMD Ryzen9 7940HX和16 GB运行内存。

本文主要基于图Transformer实现基于多头注意力机制的HT检测方法。模型参数见表1，其中，GNN采用的TransformerConv^[8-9]是一个基于多头注意力机制的Transformer操作层，用于实现图神经网络中的消息传递。

表1 模型参数

架构参数		训练参数	
输入层尺寸	动态	优化器	AdamW
隐藏层尺寸	64	学习率	0.005
图神经网络层数/层	3	权重衰减	1×10^{-4}
注意力头数/个	2/4/6/8	运行次数/次	200
随机失活概率	30%	损失函数	Focal Loss、BCE
激活函数	ReLU	域分类权重损失	0.3

3.1.2 数据集设置

本文实验使用Trust-Hub公共存储库中3组不同标准单元库的网表，实验数据集见表2。第一组为基于6个基准电路的门级网表库SAED，该网表库被领域内大多数研究所使用，第二组为TRIT-TC数据集，其在8个大规模的基准电路集中插入组合特洛伊木马，第三组为TRIT-TS数据集，其在4个大规模的基准电路集中插入触发的FSM和计数器特洛伊木马。

表2 实验数据集

数据库	网表数量	网表信息
SAED	21个	在6个不同电路设计中随机插入组合逻辑HT和时序逻辑HT
TRIT-TC	580个	在8个不同电路设计中随机插入 n 触发组合逻辑HT
TRIT-TS	334个	在4个不同电路设计中随机插入FSM和计数器类型时序逻辑HT

3.1.3 评价指标

基于HT检测任务中通常关注的3项关键特性：极端类别不平衡、高漏检风险代价和高误报复查成本，本文使用该类任务中常用的F1分数、召回率、精确率、准确率和真阴率来评估模型的检测能力，实验指标见表3。其中，TN为正常节点被预测为正常节点的数量，TP为木马节点被预测为木马节点的数量，FN为木马节点被预测为正常节点的数量，FP为正常节点被预测为木马节点的数量。

表3 实验指标

实验指标	含义	计算公式
准确率 (Accuracy)	所有节点中被正确预测的节点所占的比例	$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$
召回率 (Recall)	木马节点被正确预测为木马节点所占的比例	$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$
真阴率 (TNR)	非木马节点中被正确预测为非木马节点所占的比例	$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}}$
精确率 (Precision)	预测为木马的节点中真正木马节点所占的比例	$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$
F1分数 (F1-Score)	精确率与召回率的调和平均数	$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

3.2 参数寻优实验

SAED数据集基本信息见表4，由于SAED数据集较小，本文实验采用留一交叉验证法对模型进行训练，该方法将一个基准电路用于测试，并在其余基准电路上训练模型，该设置同时满足无黄金电路参考和HT未知的检测要求。

类似于卷积神经网络（CNN）中的卷积核大小和步长决定其卷积层的感受野和滑动距离，图Transformer模型的注意力头数决定了模型中并行的注意力机制数量。因此，本文实验主要探究图Transformer层中注意力头数（2/4/6/8个）以及融合方式（拼接/平均）对木马检测性能的影响。为保证公平比较，训练中超参数设置和训练轮数均保持一致。不同融合方式实验结果对比如图4所示。

表4 SAED数据集基本信息

电路名称	正常节点数量/个	木马节点数量/个	触发器	作用类型
RS232-T1000	223	13	组合比较器	功能篡改
RS232-T1100	224	12	顺序比较器	功能篡改
RS232-T1200	223	14	顺序比较器	功能篡改
RS232-T1300	224	9	组合比较器	功能篡改
RS232-T1400	223	13	顺序比较器	功能篡改
RS232-T1500	226	14	顺序比较器	功能篡改
RS232-T1600	224	12	顺序比较器	功能篡改
S15850-T100	2 106	27	组合比较器	拒绝服务攻击
S35932-T100	5 427	15	比较器	拒绝服务攻击
S35932-T200	5 426	16	比较器	拒绝服务攻击
S35932-T300	5 447	36	比较器	拒绝服务攻击
S38417-T100	5 328	12	比较器	拒绝服务攻击
S38417-T200	5 329	15	比较器	拒绝服务攻击
S38417-T300	5 331	15	比较器	拒绝服务攻击
S38584-T100	6 473	9	比较器	拒绝服务攻击
S38584-T200	6 371	83	状态机	信息泄露
S38584-T300	6 473	731	状态机	信息泄露
EthernetMAC10GE_T700	102 492	13	顺序比较器	功能篡改
EthernetMAC10GE_T710	102 492	13	顺序比较器	功能篡改
EthernetMAC10GE_T720	102 492	13	顺序比较器	功能篡改
EthernetMAC10GE_T730	102 492	13	顺序比较器	功能篡改

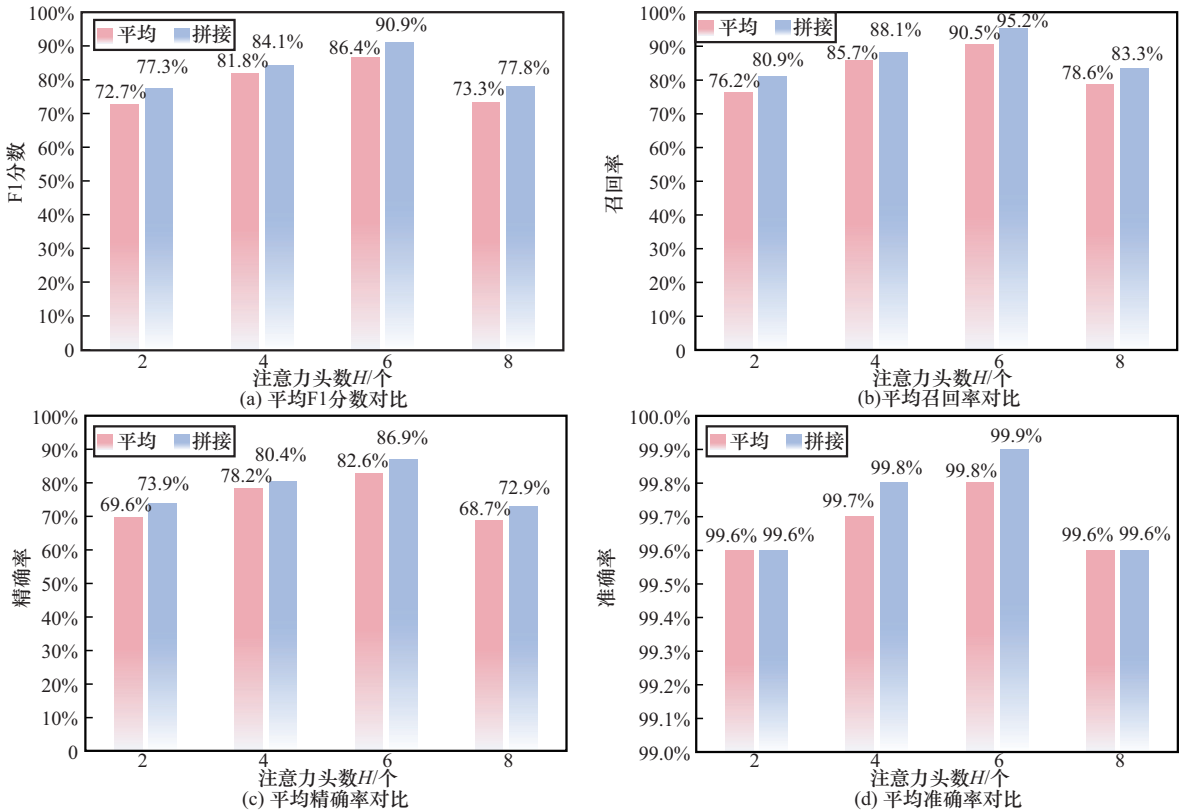


图4 不同融合方式实验结果对比

由图4可以看出,随着注意力头数 H 增加,F1分数、召回率和精确率先上升后趋于稳定,注意力头数为6个时检测达到最佳性能。此外,特征融合采用拼接方式的各指标结果均优于采用平均方式。综上所述,多头注意力机制能有效捕捉不同层次的图结构特征。而注意力头数过多则会导致过拟合或计算冗余,性能反而下降。此外,特征融合时采用拼接方式能保留更多特征信息,适合异构性强、特征维度差异大的图数据,而采用平均方式则会丢失部分特征差异,导致信息压缩,影响模型判别能力。

3.3 对比实验

基于上述实验,采用注意力头数为6个和拼接融合的方式在SAED数据集上得到最佳检测结果,结果见表5。

表5 本文方法在SAED数据集上采用拼接融合的检测结果

电路名称	准确率	精确率	召回率	F1分数
RS232-T1000	100.0%	100.0%	100.0%	100.0%
RS232-T1100	100.0%	100.0%	100.0%	100.0%
RS232-T1200	100.0%	100.0%	100.0%	100.0%
RS232-T1300	100.0%	100.0%	100.0%	100.0%
RS232-T1400	100.0%	100.0%	100.0%	100.0%
RS232-T1500	100.0%	100.0%	100.0%	100.0%
RS232-T1600	99.7%	100.0%	88.9%	94.1%
S15850-T100	99.9%	53.5%	85.2%	65.7%
S35932-T100	99.9%	87.5%	93.3%	90.3%
S35932-T200	99.9%	66.7%	100.0%	80.0%
S35932-T300	99.9%	89.7%	97.2%	93.3%
S38417-T100	99.9%	85.7%	100.0%	92.3%
S38417-T200	99.9%	72.2%	86.7%	78.8%
S38417-T300	99.6%	72.0%	85.7%	78.3%
S38584-T100	99.9%	57.1%	88.9%	69.6%
S38584-T200	99.9%	96.4%	97.6%	97.0%
S38584-T300	99.8%	98.4%	99.7%	99.0%
EthernetMAC10GE_T700	100.0%	100.0%	100.0%	100.0%
EthernetMAC10GE_T710	99.9%	78.6%	84.6%	81.5%
EthernetMAC10GE_T720	99.9%	80.0%	92.3%	85.7%
EthernetMAC10GE_T730	100.0%	100.0%	100.0%	100.0%
平均值	99.9%	87.5%	95.2%	91.2%

由表5可知,本文提出的图Transformer模型在SAED数据集的大部分基准电路上取得较高准确率、精确率、召回率和F1分数,特别是在RS232系列和EthernetMAC10GE等大型电路上表现出优异性能。然而,在少数电路(如S15850-T100)上,精确率(53.5%)与召回率(85.2%)出现较明显不平衡,且F1分数相对较低(65.7%),表明在某些特定电路结构或木马类型中,电路中的木马触发条件极为隐蔽,与正常逻辑在结构特征上高度相似,模型对其判别边界仍存在优化空间。

各模型在SAED数据集上的测试结果对比见表6,本文方法在SAED数据集上取得了最优的综合性能(F1分数:91.2%)。在召回率方面,显著高于Netlist-GNN(84.0%)和GraphSAGE-LSTM(88.1%),高达95.2%,仅次于R-HTDetector(96.8%),基于对抗训练的R-HTDetector通过生成并学习最具攻击性的对抗样本,使分类器在特征空间中把尽可能多的真实HT网表囊括进决策边界,因此具有较高的召回率,然而受精确率影响,其F1分数仅为59.9%,导致其误判率过高。在精确率方面,本文方法为87.5%,略低于GraphSAGE-LSTM方法(94.0%),LSTM的门控机制在每个节点聚合邻居信息时,能够遗忘无关或噪声邻居特征、强化与木马行为相关的特征,从而确保分类器极少正常节点被误判为木马节点,然而,其严苛筛选同时会遗漏部分触发条件隐蔽或特征微弱的木马节点,导致其召回率仅为88.1%,存在较多遗漏木马。在综合评估指标F1分数方面,本文方法取得了91.2%的结果,在所有对比方法中表现最优,表明其在精确率与召回率之间具有更佳的平衡能力。与当前性能较优的GraphSAGE-LSTM方法相比,本文方法在F1分数上有所提高,同时在召回率方面提高了7.1%,表明本文方法对隐蔽型HT具备更强的检测能力。

综上所述,基于图注意力机制的检测模型通过动态权重聚合和全局依赖关系建模,能够显著提升对HT的特征表征能力和检测灵敏度,为解决极端不平衡条件下的硅前HT检测问题提供了一种更为有效的解决方案。下面将在大规模数据

辑木马检测中分别达到86.5%和89.2%的平均召回率，平均真阴率均稳定于99.8%，具有极低的误报水平。时序类木马检测中F1分数达到88.7%，优于组合逻辑场景（81.3%），说明模型能更有效地捕捉状态机等时序结构的异常，域不变分类机制能够有效缓解工艺差异导致的特征分布偏移，充分验证了本文方法在跨设计应用中的潜力。

泛化能力结果对比如图5所示，在TRIT-TC组合逻辑木马检测中，本文方法在召回率上达到86.5%，略低于GraphSAGE-LSTM（91.8%），但F1分数达到81.3%，优于GraphSAGE-MEAN（78.8%）和R-HTDetector（78.3%），表明其在组合逻辑场景下具备良好的检测能力。其中，由于组合逻辑木马通常依赖局部结构特征，而图Transformer的全局注意力机制在建模过程中可能引入部分冗余信息，对局部关键模式的判别能力略有下降，而GraphSAGE-LSTM通过门控机制有效筛选邻居信息，因此在组合逻辑检测中表现更优。

在TRIT-TS时序逻辑木马检测中，本文方法在召回率（89.2%）和F1分数（88.7%）上均显著优于所有对比方法，该提升主要源于时序类木马通常具有跨层级的触发路径和状态机行为，其异常模式依赖于全局结构信息，而图Transformer通过多头注意力机制能够有效建模节点间的长程依赖关系，从而更准确地捕捉时序木马中的隐蔽路径。相比之下，Netlist-GNN、Graph SAGE-

MEAN和GraphSAGE-LSTM等传统GNN方法受限于局部邻域聚合机制，难以建模长距离依赖，因此在时序木马检测中表现较弱。R-HTDetector在召回率上表现较好，但其对抗训练机制在跨域场景下容易导致误报率上升，F1分数显著下降。

3.5 消融实验

为定量评估本文所提出的图Transformer编码模块与DIC模块各自对模型性能的贡献，设计并进行了系统的消融实验。实验在SAED数据集上进行，采用与3.3节相同的留一交叉验证设置，以确保结果的可比性。本文构建了以下4个模型进行对比。本文方法：即第2节所述的完整模型。 M_1 模型：在完整模型的基础上去除DIC模块，仅使用单一任务分类器进行木马检测，不再进行域对抗训练。 M_2 模型：将完整模型中的图Transformer编码模块替换为3层GCN。 M_3 模型：将完整模型中的图Transformer编码模块替换为3层GAT。所有对比模型均保持其余模块及超参数设置与完整模型一致。

消融实验均在相同软硬件配置下独立重复5次，采取结果的算术平均值及标准差。对结果进行统计分析，其中，采用均值差异指标反映消融模型与完整模型在F1分数上的平均差距，并采用95%置信区间提供该差异的估计范围，表明在95%的置信水平下，真实性能差异有较高概率落在此区间内。

消融实验结果见表8。数据显示，完整模型

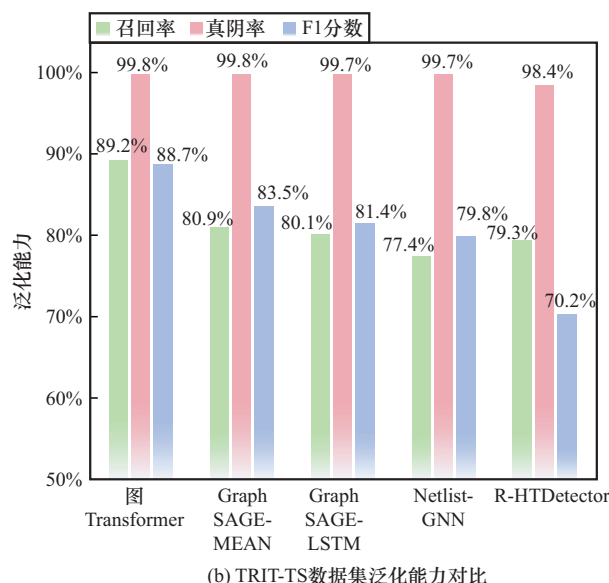
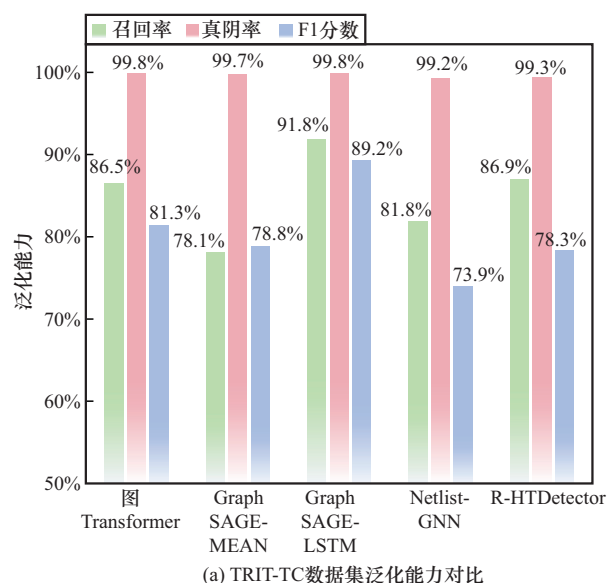


图5 泛化能力结果对比

表 8 消融实验结果

方法	准确率	精确率	召回率	F1 分数	均值差异	95% 置信区间
本文方法	$99.9\% \pm 0.01\%$	$87.5\% \pm 0.04\%$	$95.2\% \pm 0.03\%$	$91.23\% \pm 0.02\%$	—	—
M_1 模型	$99.8\% \pm 0.03\%$	$85.2\% \pm 0.05\%$	$91.8\% \pm 0.07\%$	$87.96\% \pm 0.05\%$	3.27	[3.21, 3.33]
M_2 模型	$99.8\% \pm 0.04\%$	$81.5\% \pm 0.17\%$	$89.7\% \pm 0.11\%$	$84.80\% \pm 0.15\%$	6.43	[6.28, 6.58]
M_3 模型	$99.8\% \pm 0.03\%$	$83.8\% \pm 0.09\%$	$90.8\% \pm 0.07\%$	$86.95\% \pm 0.08\%$	4.28	[4.20, 4.36]

在 F1 分数上取得了最优性能，平均 F1 分数为 ($91.23\% \pm 0.02\%$)。与之相比， M_1 模型去除 DIC 导致 F1 分数下降 3.27 个百分点 (95% 置信区间: [3.21, 3.33])，精确率与召回率亦同步降低，证实了域对抗训练对于学习域不变特征、缓解分布偏移的有效性。当使用 GCN 或 GAT 替换图 Transformer 编码模块后，模型性能出现显著下降。其中， M_2 模型的 F1 分数降至 ($84.80\% \pm 0.15$)，较完整模型下降 6.43 个百分点 (95% 置信区间: [6.28, 6.58])； M_3 模型的 F1 分数为 ($86.95\% \pm 0.08\%$)，下降 4.28 个百分点 (95% 置信区间: [4.20, 4.36])。同时，统计分析中 95% 置信区间范围集中，指向明确的性能损失，证实了差异的实质性与可靠性。综上所述，完整架构中每个模块对模型性能都产生了重要贡献，该结果充分证明了当前架构设计的合理性与完备性。

3.6 讨论

从 SAED、TRIT-TC 和 TRIT-TS 数据集的实验结果来看，本文提出的图 Transformer 检测模型在多数场景下表现出良好的检测能力。此外，通过消融实验证明了本文方法的合理性，图 Transformer 编码模块和 DIC 模块均会对实验结果产生重要影响。值得注意的是，在 TRIT-TS 时序逻辑木马检测中，模型的 F1 分数提升显著，而在 TRIT-TC 组合逻辑木马检测中提升相对较小。该现象是由于时序类木马通常具有跨层级的触发路径和状态机行为，其异常模式依赖于全局结构信息，而图 Transformer 通过多头注意力机制能够有效建模节点间的长程依赖关系，从而更准确地捕捉时序木马中的隐蔽路径。相比之下，组合逻辑木马的异常特征更多体现在局部电路结构中，全局注意力机制在建模过程中容易引入部分冗余信息或导致局部特征平滑，从而削弱了对组合逻辑关键模式的判别能力。此外，域对抗训练机制在提升跨工

艺泛化能力的同时，也在一定程度上降低了模型对局部结构差异的敏感性。因此，本文方法在处理具有复杂时序行为的木马时表现优异，而在面对局部特征敏感的组合逻辑木马时仍有优化空间。

4 结束语

本文提出了一种基于图注意力机制的门级 HT 检测方法，用于对门级 HT 进行精准检测。该模型采用图 Transformer 架构，通过多层级拓扑特征提取、图注意力编码与域对抗训练相结合，实现了在不需要黄金电路参考的条件下，有效提升跨工艺环境下对 HT 的识别能力与泛化性能。在 SAED、TRIT-TC 和 TRIT-TS 3 个基准数据集上的实验表明，本文方法在召回率、F1 分数等关键指标上优于现有主流方法，尤其在高隐蔽性时序木马检测方面表现出良好效果，为第三方 IP 核的安全验证提供了可行的技术途径。未来工作将集中于提升对极低活跃触发型木马的检测敏感度，并探索该框架在硬件安全其他任务中的应用潜力。

参考文献:

- [1] Xiao K, Forte D, Jin Y, et al. Hardware trojans: lessons learned after one decade of research[J]. ACM Transactions on Design Automation of Electronic Systems, 2017, 22(1): 1-23.
- [2] Zhang J, Yuan F, Wei L X, et al. VeriTrust: verification for hardware trust[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2015, 34(7): 1148-1161.
- [3] Lyu Y D, Mishra P. Automated trigger activation by repeated maximal clique sampling[C]//Proceedings of the 2020 25th Asia and South Pacific Design Automation Conference (ASP-DAC). Piscataway: IEEE Press, 2020: 482-487.
- [4] Elshamy M, Di Natale G, Sayed A, et al. Digital-to-analog hardware Trojan attacks[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2022, 69(2): 573-586.
- [5] Tebyanian M, Mokhtarpour A, Shafieinejad A. SC-COTD: hardware Trojan detection based on sequential/combinational testability features using ensemble classifier[J]. Journal of Electronic Testing, 2021, 37(4): 473-487.
- [6] Pan Z X, Mishra P. A survey on hardware vulnerability analysis us-

- ing machine learning[J]. IEEE Access, 2022, 10: 49508-49527.
- [7] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM Press, 2017: 6000-6010.
- [8] Li P, Wang Y, Wang H, et al. Distance encoding: design provably more powerful neural networks for graph representation learning[J]. Advances in Neural Information Processing Systems. 2020, 33: 4465-4478.
- [9] Dwivedi V P, Bresson X. A generalization of transformer networks to graphs[PP]. arXiv (2021-01-24)[2025-11-01]. arXiv: arXiv. 2012. 09699.
- [10] Shakya B, He T, Salmani H, et al. Benchmarking of hardware trojans and maliciously affected circuits[J]. Journal of Hardware and Systems Security, 2017, 1(1): 85-102.
- [11] Yu S C, Gu C Y, Liu W Q, et al. Deep learning-based hardware Trojan detection with block-based netlist information extraction[J]. IEEE Transactions on Emerging Topics in Computing, 2022, 10(4): 1837-1853.
- [12] Fyrbiak M, Wallat S, Reinhard S, et al. Graph similarity and its applications to hardware security[J]. IEEE Transactions on Computers, 2020, 69(4): 505-519.
- [13] 梅波, 郭威, 曹志鹏, 等. 基于GSCPN的HT攻击建模与安全性分析[J]. 网络与信息安全学报, 2024, 10(4): 72-84.
Mei B, Guo W, Cao Z P, et al. Modeling and security analysis of hardware Trojan attack based on GSCPN[J]. Chinese Journal of Network and Information Security, 2024, 10(4): 72-84.
- [14] 武玲娟, 朱嘉诚, 唐时博, 等. 基于可满足性无关项的HT设计与检测[J]. 网络与信息安全学报, 2021, 7(2): 35-42.
Wu L J, Zhu J C, Tang S B, et al. Design and detection of hardware Trojan based on satisfiability don't cares[J]. Chinese Journal of Network and Information Security, 2021, 7(2): 35-42.
- [15] Piccolboni L, Menon A, Pravadelli G. Efficient control-flow sub-graph matching for detecting hardware trojans in RTL models[J]. ACM Transactions on Embedded Computing Systems, 2017, 16(5): 1-19.
- [16] Chen X M, Liu Q Y, Yao S, et al. Hardware Trojan detection in third-party digital intellectual property cores by multilevel feature analysis[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2018, 37(7): 1370-1383.
- [17] Shen H H, Tan H Z, Li H W, et al. LMDet: a "naturalness" statistical method for hardware Trojan detection[J]. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2018, 26(4): 720-732.
- [18] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[PP]. arXiv (2017-02-22) [2025-11-01]. arXiv: arXiv. 1609.02907.
- [19] Veličković P, Cucurull G, Casanova A, et al. Graph attention networks [PP]. arXiv (2018-02-04) [2025-11-01]. arXiv: arXiv. 1710.10903.
- [20] Hamilton W L, Ying R, Leskovec J. Inductive representation learning on large graphs[PP]. arXiv (2018-09-10). arXiv: arXiv. 1706. 02216.
- [21] Yasaei R, Yu S Y, Al Faruque M A. GNN4TJ: graph neural networks for hardware Trojan detection at register transfer level[C]//Proceedings of the 2021 Design, Automation & Test in Europe Conference & Exhibition (DATE). Piscataway: IEEE Press, 2021: 1504-1509.
- [22] Yasaei R, Chen L K, Yu S Y, et al. Hardware Trojan detection using graph neural networks[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2025, 44(1): 25-38.
- [23] 史江义, 温聪, 刘鸿瑾, 等. 基于图神经网络的门级HT检测方法[J]. 电子与信息学报, 2023, 45(9): 3253-3262.
Shi J Y, Wen C, Liu H J, et al. Hardware Trojan detection for gate-level netlists based on graph neural network[J]. Journal of Electronics & Information Technology, 2023, 45(9): 3253-3262.
- [24] Ma P J, Li J, Liu H J, et al. Hardware Trojan detection methods for gate-level netlists based on graph neural networks[J]. IEEE Transactions on Computers, 2025, 74(5): 1470-1481.
- [25] Yasaei R, Faezi S, Al Faruque M A. Golden reference-free hardware Trojan localization using graph convolutional network[J]. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2022, 30(10): 1401-1411.
- [26] Li S, Song S J, Huang G, et al. Domain invariant and class discriminative feature learning for visual domain adaptation[J]. IEEE Transactions on Image Processing, 2018, 27(9): 4260-4273.
- [27] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2017: 2999-3007.
- [28] Ganin Y, Lempitsky V. Unsupervised domain adaptation by back-propagation[PP]. arXiv (2015-02-27) [2025-11-01]. arXiv: arXiv. 1409. 7495.
- [29] Hasegawa K, Hidano S, Nozawa K, et al. R-HTDetector: robust hardware-Trojan detection based on adversarial training[J]. IEEE Transactions on Computers, 2023, 72(2): 333-345.

[作者简介]



徐宇航 (2002-), 男, 信息工程大学硕士生, 主要研究方向为HT防御和内生安全。



于洪 (1988-), 女, 博士, 信息工程大学副研究员, 主要研究方向为网络空间内生安全和拟态防御。



郑锐 (1999-), 男, 信息工程大学硕士生, 主要研究方向为芯片设计与验证和高速互联结构设计。