最近在深度学习方面的进展,尤其是在生成模型方面, 促成了高度逼真虚拟媒体的创造。Deepfake 一词指的是通 过改变或替换图像、视频或音频中某人的外貌或声音生成 的虚拟内容、这使得与真实媒体的区别变得越来越困难。 Deepfake 技术的兴起对数字媒体的准确性和可信度带来 了严重挑战,并在政治、媒体和娱乐等领域引发了关注。 在这一背景下, 最近的报告指出, 在韩国针对年轻女性的 deepfake 色情片激增,其中包括未成年人。令人担忧的是, 根据韩国教育部的数据,韩国教师工会报告称,超过 200 所学校受到了影响, 近年来针对教师的 deepfake 显著增 加。在乌克兰, deepfake 被用于在持续的冲突期间传播错 误信息和操控公众认知。这些 AI 生成的内容通过逼真地 伪造事件或声明,导致公众混淆,并使辨别真相与欺骗变 得复杂。随着 deepfake 技术的不断发展, 区分真实和虚拟 内容的能力变得越来越困难。随着 deepfake 视频和图像的 普及, 它不仅助长了错误信息的传播, 还对隐私、安全和 公众信任构成重大威胁。这些问题已驱动广泛的研究以检 测 deepfake, 并强调迫切需要有效的解决方案来应对这一 日益严重的问题。

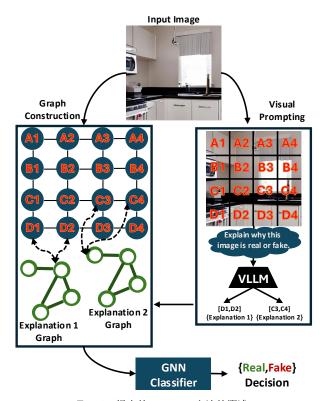


Fig. 1. 提出的 ViGText 方法的概述。

现有方法及其局限性。最近在检测深伪图像方面的努力主要依赖于基于学习的方法。这些方法通常从包含真实和深伪图像的标记数据集开始,目标是训练一个模型来检测深伪图像,并将这种检测能力推广到新的、未预见的深伪图像中。常用的模型包括卷积神经网络(CNNs)以及更简单、更传统的模型,如前馈神经网络。[?] 使用 CNN 进行图像块分类,而 [?] 则通过 Gram-Net 捕捉全局纹理来增强检测能力。[?] 着重于重新合成图像以分析残差错误,[?] 在 CNN 分类器中展示出跨体系结构的泛化能力。此外,[?] 介绍了针对介观图像特性的轻量级 CNN。最近,已经

转向更简单的模型和新颖的特征提取技术。[?] 使用频域特征与基本的前馈神经网络,[?] 结合了来自基础模型的图像和字幕特征,[?] 表明在一个非深伪专用的大型基础模型上训练的线性层可以实现有效的分类和泛化。Sifat 等人在[?] 中分析了当前的深伪检测方法并确定了关键挑战。他们证明了许多现有的方法在暴露于用户定制或微调版本的生成模型时,不能有效泛化。此外,他们指出这些方法容易受到使用先进基础模型生成的对抗攻击。这些发现揭示了当前技术中的一个重要缺口,即在适应现实世界情景中深伪的不可预测性方面存在困难。

挑战。深伪检测面临重大挑战,特别是在应对不断演变的威胁环境方面实现稳健性。生成性人工智能技术的快速发展经常超过检测能力,使得防御措施难以准备应对新的攻击向量 [?]。另一个关键挑战在于文本和视觉数据的整合。当前的方法依赖于图像标题,这些标题缺乏捕捉有效检测所需的细微特征的深度和特异性 [?]。即使在详细文本数据可用的情况下,将这些信息与视觉数据结合起来也并非易事 [?]。传统方法通常以简单的方式连接视觉和文本嵌入,未能充分利用它们的互补特性。另一个挑战是需要在各种深伪模型上实现广泛的通用性。当应用于微调或定制的生成模型时,现有方法往往表现出显著的性能下降,这强调了它们对新威胁的适应性不足 [?]。为克服这些挑战,探索能够有效结合文本和视觉信息的高级整合策略,并整合它们的互补优势是至关重要的。

提出的 ViGText 概述. 为了克服上述挑战, 我们提出了 ViGText ,这是一种新的深度伪造检测方法,它在图框架 中集成了来自视觉大型语言模型(VLLM)的图像分析和 基于文本的解释,如图 1 所示。ViGText 的创新在于其能 够通过量身定制的图形构建统一视觉和文本分析,从而能 够以增强的泛化和鲁棒性检测出细微的不一致性。该过程 首先将输入图像划分为多个方形补丁,每个补丁在图像图 中表征为一个节点。每个补丁的嵌入都会考虑空间和频率 信息,后者通过离散余弦变换(DCT)提取。然后添加边 以连接相邻的补丁,以捕捉局部空间依赖关系。除了图像 图之外,还使用 VLLM 为补丁生成文本解释构建解释图。 解释图与图像图集成,使每个解释与其描述的补丁连接, 从而形成双图结构。然后通过图神经网络(GNN)分析该 双图,以确定图像是真实的还是伪造的。虽然最近的方法 使用图像标题进行深度伪造检测 [?] , ViGText 在图结构 中集成详细的文本解释、结合空间和频率特征、以提高对 微调模型的泛化以及对对抗性图像的鲁棒性。

贡献总结。该工作提出了以下贡献。

- 用于增强检测的双图框架介绍: 我们提出了 ViGText , 这是一种将图像分析与 VLLMs 生成的文本解释相统一的 新方法。ViGText 在考虑空间和频率特征的同时嵌入每个 图像补丁,然后将视觉和文本数据组织成一个双图结构,从而实现更稳健的深度伪造检测集成。
- 增强对多样化生成模型的泛化能力: ViGText 在无需训练基础模型图像的情况下,实现了对多种用户定制化、微调的生成模型变体的卓越泛化能力。通过在图框架中有效整合上下文感知解释和频域特征,ViGText 缓解了以往方法在面对用户定制化模型时出现的显著性能下降。
- 针对不断演变的威胁的稳健性: ViGText 在对抗攻击中表现出强大的抗逆性,包括基于新型基础模型的威胁。这种稳健性解决了现有方法在面对通过先进的视觉基础模型刻意打造的对抗性操控时失败的问题。

• 在泛化数据集上进行扩展测试: 为了在多样化的场景中评估泛化能力, 我们引入了一个扩展数据集, 该数据集包括八个从用户自定义微调的 Stable Diffusion 3.5 模型 [?] 的变体中派生出的新的测试集。这一扩展结合了现有的数据集, 使得能够更全面地评估在更广泛的生成模型范围内的检测性能。

## I. 背景

深度伪造。深度伪造是一种合成媒体形式,使用人工智能(AI)创建大规模仿真但虚假的图像、视频或音频录制。深度伪造技术涉及先进的机器学习技术,特别是深度学习算法,它们分析大量真实图像或音频数据集以生成新的、高度逼真的内容[?],[?]。最初开发用于娱乐和创意目的,深度伪造迅速发展,引发了重大伦理、法律和社会关注[?],[?]。深度伪造可以被恶意使用来伪造个人的影像,展示其从未说过或做过的事情,导致潜在的伤害,如误导信息、身份盗用和声誉损害[?],[?]。深度伪造技术日益增加的可获取性引发了全球对于需要进行监管、检测方法以及提高公众意识的辩论,以减轻这种强大技术带来的风险。

图神经网络。图神经网络(GNNs)[?] 通过将深度学习扩展到图结构数据,显著推动了该领域的发展。这些模型在图的边上传递信息,并在节点处汇总这些信息。GNN的工作流程包括利用局部节点特征和图拓扑从图输入中提取低维嵌入。GNN在各种领域中被认为是有效的分类器。例如,在欺诈检测中[?],它们利用交易网络中的关系信息。同样,在药物发现中[?],GNN有助于理解分子结构。它们还成功应用于社交网络分析[?]和推荐系统[?],展示了其作为分类器的多样性和强大性。GNN处理复杂关系数据的能力以及其适应不同类型图结构信息的灵活性,使其成为众多最新应用中的强大工具。

自然语言处理中的图神经网络图神经网络(GNNs)通过图表示有效地捕获文本中的复杂依赖关系,成为自然语言处理(NLP)中的一种强大工具。与传统的基于序列的模型不同,GNNs 能够通过将单词、句子或文档表示为图中的节点,并用边来表征各种语言连接,从而对句法和语义关系进行建模[?],[?]。这种方法在关系抽取[?]、文本分类[?],[?] 和情感分析[?]等任务中特别有效,因为这些任务中理解语言的内在结构至关重要。

视觉大型语言模型和视觉提示视觉大型语言模型 (VLLMs) 是一个快速发展的人工智能领域,它结合了视觉和文本数据来执行广泛的任务,包括图像描述 [?],[?]、问题回答 [?]、分类和分割 [?]。该领域最近一个显著的发展是视觉提示的概念 [?],[?]。这种技术涉及使用特定的视觉指令来指导模型对文本的解释或生成,类似于在自然语言处理中使用文本提示来根据文本提示生成响应。在视觉提示中,模型被给予一个图像或图像的修改版本,而不是仅仅依赖于文本输入。最近的研究已证明视觉提示在提高 AI 模型在各种任务中的适应性和性能方面的有效性。通过结合标记或注释等视觉指令,这些技术帮助模型应对复杂任务,如机器人操作 [?]、图像处理 [?] 和感知 [?],而无需额外的微调。

# II. 威胁模型和假设

威胁模型包括一个生成深度伪造并以规避检测为目标的 对手(恶意行动者),而防御者尝试检测这些深度伪造。这 种互动在图 2 中有所描绘。我们通过详细描述对手和防御

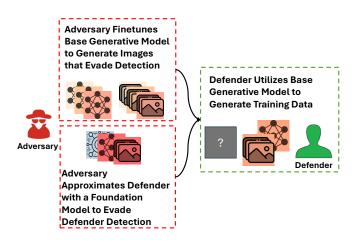


Fig. 2. 威胁模型的示例图解。

者的目标、知识和能力来刻画威胁模型。对手的目标是产 生能规避防御者系统检测的深度伪造。对手具备先进的生 成技术,包括微调基础模型以创造多样且逼真的深度伪造 的能力。近来的参数高效微调技术, 例如低秩适配 (LoRA) [?] ,使得即便是计算资源有限的对手也能创建基础模型 (如 Stable Diffusion [?] ) 的定制化变体。 这些微调变体引 人的细微变化使得检测变得具有挑战性,甚至在没有防御 者方法的明确知识时也会降低检测性能 [?] 。此外,尽管 我们的实验涉及使用 StyleCLIP 进行的编辑,这是一种在 StyleGAN2 [?] 潜在空间中操作的文本驱动图像编辑方法, 但我们并不声称这些代表现实中的部分操控,如面部交换 或重演。相反,我们使用这些编辑来模拟一种独特的对抗 行为:最小且局部的语义变化,保留身份和整体图像上下 文。这属于更广泛的全合成图像类别,但为研究通过细微 而非剧烈变化来规避检测的目标性操控提供了一个有用的 起点。这些受控操控帮助我们研究检测小幅但意义重大的 变化的挑战,为我们的威胁模型添加了一种不同类型的对 抗策略。对于防御者来说,其目标是高效地检测深度伪造, 并在广泛的微调变体中推广检测能力。访问对手用来派生 其微调变体的基础生成模型是一种实际且有效的方法,因 为许多生成模型是公开可用的或可以在像 huggingface 这 样的网页上广泛访问。像 LoRA 这样的微调技术通常仅修 改特定的层或参数,同时保留基础模型的核心特征。这些 共享特征,比如架构模式、特征表示和生成倾向,在变体 中基本保持完整。通过专注于这些基础特征,防御者可以 将检测能力推广到微调变体,而不需要进行每一个可能的 变体的训练,这是不切实际的。防御者的另一个目标是确 保对抗对抗攻击时的鲁棒性、特别是那些使用先进基础模 型生成的攻击。这些模型使得对手能够制造微妙且高度具 有欺骗性的操控,从而绕过检测。鲁棒性对于在面对愈加 复杂和适应性的威胁时,维持检测系统的可靠性至关重要。 为了实现这个目标,防御者旨在创建能够抵御不断变化的 攻击策略并保持结果完整性的系统。

# III. 建议的 VIGTEXT

在本节中,我们详细介绍了 ViGText 方法,该方法在图??的框图中进行了说明,涉及从图像块和生成的解释中构建图,提取特征,并在使用 GNN 进行检测之前整合这些图。我们首先讨论使用 VLLMs 生成的解释的动机。在此

之后,我们构建了具有文本图像的深度伪造检测问题。然后,我们引入了一个基于图的框架,该框架结合了文本和图像数据,为分析提供了更丰富的上下文。最后,我们描述了生成这些文本解释的过程及其与图像的整合,形成统一的图结构。

## A. 从标题到解释



Fig. 3. 生成的图像被 DE-FAKE [?] 错误分类为真实图像。

ViGText 基于现有技术使用的类似概念,如 DE-FAKE [?] ,它将图像标题与视觉数据相结合以进行深度伪造检测。然而,标题通常仅提供图像的广泛描述,缺乏识别不一致性所需的特定性。例如,如图 3 所示,DE-FAKE 根据标题"厨房和餐饮区"误将深度伪造图像分类为真实,因为该标题用通用术语描述场景,没有涉及可能指示操控的视觉细节。例如,VLLM 生成的解释可能说,"橱柜和悬挂灯显示自然反射和阴影,表明这是一个真实环境",或"桌子和椅子具有细致的纹理和一致的光照,这是真实图像的特点"。

这些详细的解释捕捉了那些对图像的真实性感知有贡献的具体特征。然而,仅靠 VLLMs 是无法准确区分图像是真实的还是伪造的 [?]。这就是 ViGText 擅长的地方。通过分析视觉内容和相应的解释,ViGText 识别出描述特征与实际视觉元素之间的不一致性。例如,如果一个解释提到了真实的阴影和反射,但图像却缺少这些元素或者显示出不自然的伪影,这种差异就是该图像可能是深度伪造的强烈迹象。将详细的解释与视觉分析相结合,使得 ViGText 能够解决诸如 DE-FAKE 等基于文字描述的方法的局限性,并提供一个更坚实和可靠的深度伪造检测框架。

# B. 问题表述

根据上述解释,深度伪造检测的问题,重点在于泛化能力和对抗性鲁棒性,需要优化一个分类器函数 f ,该函数将输入图像 I 及其对应的解释 E 映射到二值输出  $\{0,1\}$  ,其中 0 表示真实图像,1 表示伪造图像。在基于机器学习

的解决方案中,目标是在数据集  $\mathcal{D} = \{(I_i, E_i, y_i)\}_{i=1}^n$  上最大化此分类器的准确性,  $y_i$  是(1)中所表达的真实标签。

Maximize: 
$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\{(I_i, E_i) = y_i)$$
Subject to: 
$$\min_{\delta \in \Delta} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(f(I_i + \delta, E_i) = y_i) \ge \tau_r,$$

$$\mathbb{E}_{(I, E, y) \sim \mathcal{D}_{new}}[\mathbb{I}(f(I, E) = y)] \ge \tau_g,$$

$$(1)$$

其中  $I_i$  是输入图像, $E_i$  是生成的解释, $y_i$  是标签(0为真实,1为伪造)。指示函数  $\mathbb{I}(\cdot)$  在条件为真时返回 1,否则返回 0。 $\Delta$  表示允许的扰动以测试鲁棒性, $\tau_r$  为所需的鲁棒性阈值。 $\mathcal{D}_{new}$  表示未见数据的分布, $\tau_g$  是必须满足的泛化阈值, $\mathbb{E}$  表示期望值。这个问题需要仔细考虑如何整合图像和文本数据,以充分理解包括文本信息的好处。f 是表征检测的二值分类器,g 是真实标签(0为真实,1为伪造), $\mathcal{L}$  表示用于评估预测准确性的 0–1 损失。

## C. 视觉与文本整合

文本和视觉数据的整合对于有效的深伪检测至关重要。 虽然 DE-FAKE [?] 使用简单的嵌入拼接来结合标题和图 像,但这种方法未能捕获详细的相互依赖性。相比之下, ViGText 使用基于图的模型,该模型整合文本解释和视觉 数据来建立有意义的关系。

为了探讨整合方法的影响,进行了以下实验,我们比较了两种方法: DE-FAKE, 它使用解释作为文本输入但将它们的嵌入与图像嵌入任意拼接, 以及 ViGText 。表 I 总结了这种比较的结果。

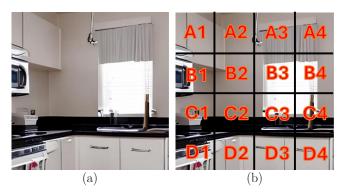


Fig. 4. 图像叠加网格: (a) 原始, (b) 带有网格叠加。

TABLE I 当给予相同的解释作为文本输入时, VIGTEXT 与 DE-FAKE [?] 的性能 比较(最高者用粗体显示)。

	Accuracy	Recall	Precision	F1
DE-FAKE w/Explanations	90.00	91.20	89.00	90.10
ViGText	99.25	99.80	98.52	99.26

表中的结果显示出明显的性能差异。简单地用解释替换标题并将其嵌入与图像特征串联,并不能导致准确的检测。尽管使用了更丰富的解释而不是简单的标题,DE-FAKE仍然实现了90%的准确率。这展示了一个关键的限制:通过串联随意结合文本和视觉数据未能捕捉到它们之间复

杂的相互依赖性。不像 DE-FAKE 中使用的简单技术,图 结构允许 ViGText 捕捉文本解释和视觉特征之间复杂的相互依赖性, 这使得理解更加深入并显著提高了检测性能。

# A Sample Textual Explanation

{ B3,B4 }: The window blinds have uneven spacing, and the light passing through does not align properly with the individual slats, which suggests an error in rendering light and shadows. { D1,D2 }: The oven appears to have a distorted handle, and the reflection and shadow around it don't conform to the expected perspective and lighting. { D3 }: The drawer underneath the stove has irregular handles that are asymmetrical, which is not typical for kitchen design and could be an oversight by the AI.

Fig. 5. 文本解释的一个例子,每个解释对应图像中的特定区域。

# D. 解释补丁集成图构建

解释生成。使用 VLLM 来生成图像是真实的还是伪造的解释是通过一种称为视觉提示的方法实现的。在 ViGText中,这涉及用一组等大小的正方形贴片覆盖图像,并逐一贴上标签(例如,A1, A2, A3, A4, B1等)。覆盖后的图像和原始图像都被输入到 VLLM,这使其能够生成与这些局部区域直接相关的解释。图 4 展示了这种网格覆盖。这种分割框架确保了解释与对应贴片的图像图形的准确集成。生成的示例解释展示在图 5 中,而提示模板在附录??的图??中提供。

基于网格的解释是必不可少的,因为尽管单个完整图像的解释可以描述整个场景,但它无法阐明如何将解释的各个部分与特定图像区域关联以便进行图构建。网格将每个解释连接到一个补丁,从而实现精确的跨模态边缘。但是,这带来了一个权衡:较小的补丁可以捕捉细节,但可能会丢失全局上下文,而较大的补丁则相反。ViGText通过选择中间大小的补丁来平衡这一点,并利用GNN的信息传递将局部和全局线索合并成连贯的、易于理解的人类推理。我们将在第IV节中更详细地分析这种权衡。

值得注意的是,ViGText 在其设计理念上,并不依赖于VLLM 作为独立的信任锚。相反,它使用 VLLM 作为一个完全本地化、由防御者控制的组件,从而避免与外部或不透明模型相关的风险。VLLM 提供照明、几何和纹理的细粒度文本描述,这些描述在双图结构中与视觉补丁特征进行交叉验证。在训练过程中,GNN 从匹配和故意不匹配的图像-解释对中学习,使其能够检测跨模态的不一致性。这确保了 ViGText 可以通过文本和视觉线索之间的统计差异揭示操作,即使是在对抗性场景下,使得整个系统更为稳健和可靠。

图像图构建。在生成解释之后,ViGText 构建一个图,该图表示图像的片段及其相应的解释。该过程从构建图像图开始,其中每个节点表示图像的一个片段,并且与相邻片段对应的节点通过无向边相连,如图 6 (b) 左侧所示。为了表示每个片段为一个节点,ViGText 使用 ConvNeXt-Large [?] ,这是一个在 LAION-5B 数据集子集上训练的基础图像特征提取模型。该模型为每个片段提取特征嵌入。此外,右图 6 (a) 所示的 DCT 变换片段也通过相同的特征提取模型生成一个嵌入。最后,两个嵌入(图像和基于

DCT 的)进行平均,以创建一个稳健且全面的特征表示,然后将其分配给图中的相应节点。这种双域表示增强了图 捕捉空间和频域伪影的能力,这对于检测深度伪造图像中的细微操作是至关重要的。

文本图构造。为了将解释表示为图,每个句子中的词被描绘为一个节点,节点之间的边反映了词语之间的语法关系,这些关系是使用 spaCy [?] 的依存分析器提取的。这个结构展示了词语在句子中是如何交互的。这种方法不仅捕捉了词语的角色,还捕捉了它们的交互,从而得到了一个全面且结构化的解释表示。ViGText 使用 Jina [?] ,一种嵌入模型,来提取词语的特征,为每个节点分配其相应的嵌入。最后,ViGText 通过将解释图中的每个节点连接到图片图中的相应补丁节点,将这些解释图与图片图集成在一起。图 6 (b) 显示了一个与其补丁节点整合的示例解释图。

# Algorithm 1 补丁和解释集成图构建

- 1: Input: An image I, image feature extraction model M, word feature extraction model B, VLLM.
- 2: Output: Patch and Explanation Word Correspondence Graph
- 3: Overlay I with the grid mask to produce the image Q.
- Query the VLLM with Q to produce patch-specific explanations.
- 5: Split I into patches and extract spatial features for each patch using M.
- 6: Apply the DCT transformation to each patch and extract frequency-domain features using M.
- 7: Average the spatial and frequency-domain features to create combined embeddings for each patch.
- 8: Construct the image graph with nodes representing patches and features corresponding to their combined embeddings.
- 9: Construct a graph for each explanation with nodes representing words and edges based on grammatical relationships, extracting word features using B.
- 10: Integrate the image graph with the explanation graphs by connecting each explanation graph node to the corresponding patch node in the image graph.
- 11: Return the unified graph containing the image graph and the explanation graphs.

构建补丁和解释集成图的过程,在算法 1 详细描述并在图 1 和 ?? 中展示,从将图像 I 覆盖网格掩膜以生成修改后的图像 Q 开始(步骤 3 )。修改后的图像用于查询VLLM,生成与网格区域相关联的补丁特定的解释(步骤 4 )。图像被分割成补丁,并使用模型 M 提取每个补丁的空间特征(步骤 5 )。此外,每个补丁经过 DCT 处理,使用相同的模型在频域提取特征(步骤 6 )。空间与频域特征平均以形成每个补丁的最终特征嵌入(步骤 7 )。这些嵌入用于构建图像图,节点表示补丁,边连接相邻补丁(步骤 8 )。对于每个解释,创建一个图,节点对应于单个间,边编码其语法关系。使用词嵌入模型 B 提取单词特征(步骤 9 )。解释图随后通过基于解释和补丁之间的空间关联将单词节点连接到其对应的补丁节点,将其与图像图整合(步骤 10 )。结合视觉和文本数据的统一图作为输出返回(步骤 11 )。

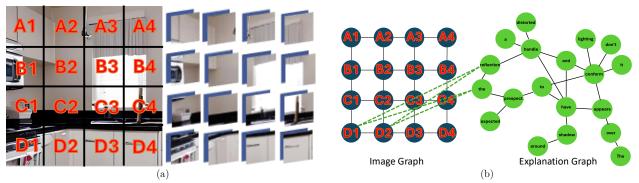


Fig. 6. 一个示例图像图及其对应的解释图构建与整合: (a) 应用网格的图像(左)以及补丁和它们对应的 DCT(右),以及(b)图像图(左)和一个示例解释图(右),仅展示连接2个节点(在实际实现中,解释图中的所有节点都连接到对应的补丁节点)。

#### IV. 实验

在本节中,我们通过一系列实验评估了 ViGText 的性能,这些实验旨在解决表格 II 中总结的问题。我们的结果表明,ViGText 在多种分类指标方面达到了最先进的检测性能。此外,ViGText 展现出强大的泛化能力,可处理来自各种微调生成模型的数据集。值得注意的是,ViGText 对基础模型驱动的对抗性攻击和以充分了解其机制为基础的定向攻击表现出鲁棒性。此外,该系统的性能对设计选择的变化依然具有弹性,表明其配置具有一定程度的灵活性。总体而言,ViGText 在保持计算成本与现有最先进方法相当或可接受的情况下,实现了这些进步。有关本研究中使用的数据集的源代码和信息,可以在以下匿名链接中找到: ViGText 。

TABLE II 研究问题和关键答案总结。

Q	Property Investigated	Key Result
1	Detection effectiveness	Highly effective
2	Generalization	Strong generalization
3	Robustness	High robustness
4	Sensitivity to design choices	Generally insensitive
5	Empirical costs	Tolerable

# A. 设置、数据集和基准

我们使用了 Sifat 等人引入的数据集 [?],这些数据集解决了现有深度伪造研究中的关键限制。具体来说,[?]强调了许多现有数据集中对内容和图像质量缺乏控制的问题,这可能导致对最先进的检测方法性能的过高估计。为了解决这个问题,[?]构建了两个精心策划的数据集,旨在改善控制,并使深度伪造检测方法的评估更加准确。

● 稳定扩散 (SD) 数据集包含来自 LAION-AESTHETICS 数据集 [?] 的真实图像和使用 Realistic Vision v1.4 模型 [?] 生成的假图像。该数据集涵盖了各种类型的内容,包括人物、自然、物体、插图和数字艺术。其结构确保了平衡,提供 16,000 张用于训练的图像,2,000 张用于验证,和 2,000 张用于测试,在真实和假图像之间均匀分配。该数据集的一个关键重点是评估深度伪造检测方法对由生成模型的微调变体生成的图像进行概括的能力,因为概括仍然是现有方法的一个持续挑战。为评估这一点,该数据集包括 16 个额外的测试集,这些测试集源自基础 SD 1.5 模型。其中,8 个测试集的图像是使用全

模型(FM)微调方法生成的— -这种方法更新所有参数 其余 8 个则是使用低阶适配(LoRA)微调 [?] 生成 的,这是一种仅更新模型部分参数的计算效率高的方法。 值得注意的是,即使没有特别的对抗意图,这些微调变体 生成的图像在许多最先进的检测方法中也导致了显著的性 能下降,如[?]所示。这凸显了微调技术普及所带来的日 益严重的威胁, 以及对强大、可推广的检测系统的迫切需 求。此外,我们通过创建8个额外的测试集来扩展泛化测 试,这些测试集对应于当前最先进的开源生成模型 Stable Diffusion 3.5 模型 [?] 的 8 个新的 LoRA 微调变体。由 于 Stable Diffusion 3.5 模型的规模显著较大,达 80 亿参 数,在此扩展中我们选择仅包括 LoRA 微调变体。虽然 LoRA [?] 微调在计算上效率较高, 但对如此大规模模型的 所有参数进行微调将需要大量计算资源,这对于大多数用 例是不切实际的。附录 VII 中的表 VIII 包含有关 Stable Diffusion 3.5 LoRA 模型 [?] 的更多信息。这一扩展在生 成 AI 最新进展所带来的挑战性场景下,对检测方法进行 了全面评估,进一步强调了检测系统需要既稳健又具有适 应性的必要性。

• StyleCLIP 数据集:该数据集包含了一组平衡的真实 和虚假人脸图像,专门设计用于使用视觉基础模型研究 对抗性攻击的鲁棒性。真实图像来源于 Flickr-Faces-HQ (FFHQ) 数据集 [?], 这是一个高质量的人脸图像集合, 而 虚假图像则是使用 StyleGAN2 [?] 生成的,这是一个广泛 采用的人脸合成生成模型。该数据集是平衡的,包括16,000 张用于训练的图像,以及各 2,000 张用于验证和测试的图 像。与 SD 数据集不同, StyleCLIP 数据集强调在对抗场 景下的鲁棒性评估。如 [?] 所详述,对抗性攻击在此情境 中涉及操控人脸图像的语义属性,如改变表情、添加配饰 或修改面部特征,而不引入可察觉的噪声。这些攻击使用 视觉基础模型作为代理来优化操控,能够生成规避检测的 对抗性深度伪造。为了彻底评估这一挑战,使用了三种先 进的基础模型, EfficientNet [?] 、ViT [?] 和 CLIPResNet [?] 作为代理模型。这些模型用于创建三个额外的对抗性 测试集,每个测试集均量身定制以利用现有深度伪造检测 方法的弱点。因此,该数据集为评估检测系统在对抗性环 境中的弹性提供了关键基准,突出了基础模型驱动攻击所 暴露的漏洞。为了扩展我们的评估,我们实施了一种更高 级的对抗性攻击,该攻击模拟了一名对 ViGText 了解甚多 的攻击者。假设这一虚构的攻击者拥有对 ViGText 使用 的训练数据集和图创建流程的详细了解。基于这些假设, 攻击者创建了一个替代模型,旨在模仿 ViGText 的功能。 替代模型包含两个图卷积层,并使用 Dinov2 [?] 和 Jina [?] 基础模型分别提取图像和文字嵌入。这一替代模型在 StyleCLIP 数据集上进行了训练,在分类指标方面取得了 高性能,准确率、召回率、精确率和 F1 分数均超过 95 %。 利用这个替代模型,我们进一步训练 StyleGAN2 [?] 生成 模型以生成对抗性图像。这些图像专门设计用来逃避替代 模型的检测,因此也逃避 ViGText 的检测。攻击通过最小 化替代模型的 logits z 与目标标签 y 之间的交叉熵损失来 优化生成器(将目标标签设为真实图像的标签以创建逃避 性图像):

$$\mathcal{L}_{\text{adv}} = -\mathbb{E}_{\boldsymbol{x} \sim G(\boldsymbol{\theta})} \left[ y \log p(\boldsymbol{z}) + (1 - y) \log(1 - p(\boldsymbol{z})) \right], (2)$$

其中  $G(\theta)$  是带参数  $\theta$  的 StyleGAN2 生成器,p(z) 表示替代模型对真实标签的输出概率,而  $x \sim G(\theta)$  是生成的图像。这个设置使我们能够评估 ViGText 对由模仿一名知识渊博的真实世界威胁攻击者制作的对抗性图像的抵御能力。

值得注意的是,虽然 StyleCLIP 数据集及其基于代理的攻击可以有效测试在强视觉操作下的鲁棒性,但它们没有考虑到同时针对图像和 VLLM 生成的解释的协调攻击。执行此类双目标攻击需要拥有对图像生成器和 VLLM 的白盒访问,以便在各个组件之间进行基于梯度的优化。这大大增加了复杂性和计算成本,使其超出了我们当前评估的范围,并成为未来研究的一个开放方向。

我们报告了本节中进行的所有实验的经典分类指标,包括准确率、精确率、召回率和 f1 分数。至于基线,我们选择了在 [?] 的分析中表现良好的三种最先进的方法。这些方法是:

- 离散余弦变换(DCT)[?]:这种方法通过使用离散 余弦变换(DCT)从图像中提取频域特征,以识别细微的 伪迹。这些特征被对数缩放以提高性能,然后用于训练一 个逻辑回归分类器,该分类器能够有效地区分真假图像。
- DE-FAKE [?]: 该方法通过使用 CLIP 模型 [?] 构建一个 deepfake 检测器,具体做法是将图像的嵌入与用于生成图像的文本提示的嵌入进行增强。这些增强后的嵌入被用于训练一个两层多层感知器作为分类器。
- UnivCLIP [?]: 这种最新的方法利用了一个大型基础模型,具体来说是 CLIP:ViT-L/14 模型 [?]。这种方法从冻结的 CLIP:ViT 模型中提取特征,然后使用最近邻分类器或线性分类层,在进一步训练的基础上,来判断图像是真实的还是伪造的。这里更倾向于使用线性分类器以获得更好的性能。

尽管最近的方法,如 ObjectFormer [?] 和基于大规模 视觉-语言模型 (VLMs) [?], [?] 的检测器表现出有希望的 检测能力,但它们也依赖于资源密集型架构,如密集的 transformer 注意力或数十亿参数的语言模型。这些设计 面临可扩展性限制,阻碍其广泛部署的实用性。相比之下,我们关注的方法是在强检测性能与计算效率之间取得平衡。ViGText 在反映现代生成技术的最新挑战性数据集上展示了最先进的准确性和鲁棒性,同时保持轻量级的基于图的架构,确保了更广泛的适用性和可扩展性。

对于 ViGText ,我们在所有实验中使用一致的 GNN 架构。该模型由三个带有两个注意力头的图注意力网络 (GAT) [?] 层组成,每层之后进行批处理归一化和 ReLU

激活。在每层之后应用 Dropout 以通过正则化训练过程来 防止过拟合。在传递到最终分类的全连接层之前,节点特 征通过全局均值池化进行聚合。

该模型使用 Adam 优化器 [?] 进行了 40 个时期的训练,以最小化交叉熵损失,并在训练过程中采用学习率调度动态调整学习率。所有实验均采用 4x4 补丁大小,并使用Qwen2-VL-7B-Instruct [?] 作为解释生成的 VLLM。实验在配备 64 GB RAM、8 GB RTX 2070 GPU 和 32 核 Intel Xeon 处理器的工作站上进行。

# B. 性能分析

TABLE III 对数据集各自测试集的性能分析(最高的以粗体显示)。

Approach		SD			StyleCLIP				
Approach	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	
DCT	85.50	83.30	88.80	85.96	98.80	98.22	99.40	98.80	
DE-FAKE	92.45	91.17	94.00	92.5	74.05	75.34	71.50	73.37	
UnivCLIP	93.04	92.33	93.89	93.10	93.04	93.79	92.19	92.99	
ViGText	99.25	99.8	98.52	99.26	99.60	99.90	99.21	99.60	

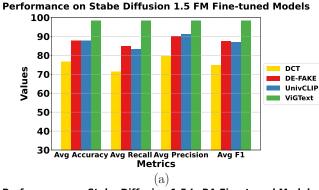
我们首先讨论问题 Q1: ViGText 在检测深度伪造方面有多有效?为评估这一点,我们将 ViGText 的性能与上述基线进行比较,使用两组数据集中的指定质量指标。正如表 III 所示,ViGText 一贯优于最新的最先进技术,展示了检测由多种方法生成的深度伪造的强大能力,并反映了深度伪造技术的实际威胁态势。

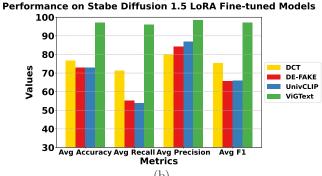
这些结果突出显示了 ViGText 通过双图结构独特整合 视觉和文本信息的有效性。通过使用空间和频率嵌入以及详细的上下文感知文本说明,ViGText 实现了优越的检测性能。附录 VIII 中报告了对来自最先进扩散 API 的图像以及在高级对抗攻击下的附加实验,并在附录 IX 中展示了仅由 ViGText 检测到的其他示例案例。

# C. 泛化

在这里,我们讨论问题 Q2: ViGText 能否很好地泛化以检测由各种微调变体生成的图像?该评估主要关注 SD 数据集。ViGText 使用该数据集的训练数据进行训练,并在 24 个单独的测试集上进行测试,这些测试集对应于不同 SD 模型的微调变体。这些测试集分为全面模型(FM)微调和 LoRA 微调变体。图 7 (a)显示了 SD 1.5 模型的8 个 FM 微调变体的平均性能指标,而图 7 (b)和图 7 (c)分别展示了 SD 1.5 和 3.5 的 16 个 LoRA 微调变体的结果。

如图 7 所示,ViGText 在所有指标上相比于基线方法表现出优越的泛化性能,这展示了它检测由不同微调模型生成的假图像的能力。鉴于传统的数据驱动方法在测试分布改变的数据(如那些来自微调生成模型的数据)时经常表现出性能下降,这一点尤其重要。相比之下,ViGText 通过其基于图的框架缓解了这一限制,该框架专注于学习从数据点派生的图的结构拓扑。此外,ViGText 使用与内容无关的频域特征,捕捉对图像内容不变的微妙特征。这些特征在增强泛化方面起着关键作用,因为它们减少了对底层数据分布或训练中使用的生成模型特征的依赖。ViGText将其基于图的架构与频域分析相结合,有效地适应多变且具有挑战性的测试场景,从而加强泛化能力。





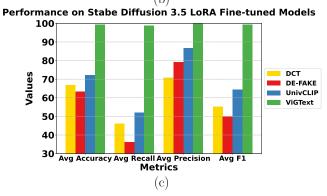


Fig. 7. 概括在以下方面的表现: (a) Stable Diffusion 1.5 FM 微调模型, (b) Stable Diffusion 1.5 LoRA 微调模型以及 (c) Stable Diffusion 3.5 LoRA 微调模型。

### D. 鲁棒性

在本小节中,我们探讨问题 Q3: ViGText 在面对基于基础模型的对抗攻击和图像操控时有多强的鲁棒性?为此,我们集中研究 StyleCLIP 数据集及其相关的对抗性操控测试集。这些测试集采用了最先进的对抗攻击 [?] 制成,该攻击利用在大量数据集上训练的基础模型生成操控,从而降低传统检测方法的性能。操控旨在利用图像的细微语义特性,使攻击更为有效,同时不引入可察觉的噪点。

我们在 StyleCLIP 数据集的训练部分上训练了 ViGText,然后在这些被操控的集合上进行测试,每个测试集对应一个不同的替代基础模型,包括 EfficientNet、ViT 和用于生成对抗攻击的 CLIP-ResNet。如图 ?? 所示, ViGText在所有评估指标上始终优于其他方法。这表明与基线方法相比, ViGText 在面对对抗性攻击时表现出显著更少的性能下降。鉴于这些攻击的复杂性,它们利用在数百万种不同图像上训练的基础模型来逼近高质量的替代模型,

ViGText 的弹性尤其值得注意。这种固有的鲁棒性使得 ViGText 成为对抗对抗性操控的有效解决方案。

接下来,我们在一个敌手对检测系统有极大了解的场景中评估 ViGText 的鲁棒性。假设敌手可以访问同样的训练数据集(StyleCLIP 数据集)以及用于构建图像和文本解释之间图结构的相同流程。这个设置模拟了一个高度有能力和见多识广的敌手,以严格评估在这种挑战性威胁模型下 ViGText 的韧性。为了模拟这个场景,我们设计了一个替代检测模型,其架构合理由地选择,模仿了 ViGText 的特征。该替代理模型使用 StyleCLIP 数据集进行训练,并在 StyleCLIP 测试集上评估时,在所有检测指标(准确率、精确率、召回率和 F1 分数)上都超过了 95 %。这一高性能的替代模型随后被用于通过优化 StyleGAN2 [?] 生成器,最小化替代模型的 logits 与目标标签(真实图像)之间的交叉熵损失,从而有效生成规避样本。

我们在这些对抗性图像上测试 ViGText , 其实现的准确率为 95.85 % , 召回率为 91.7 % , 精确率为 99.2 % , 和 F1 得分为 95.67 % , 相比之下,原始指标分别为准确率 99.6 % , 精确率 99.9 % , 召回率 99.21 % , 和 F1 得分 99.6 % 。虽然这些对抗性图像相比于基于基础模型的攻击对 ViGText 的性能影响更大,但这个结果是可以预料的,因为代理与实际检测系统非常相似,这为攻击者提供了相当大的优势。

之前的结果突出了 ViGText 的鲁棒性,即使面对对其设计和训练数据有显著了解的对手。图形化框架、频域特征的整合以及视觉和文本信息的有效结合,使得 ViGText 在这种极端威胁模型下仍能保持强大的检测性能。这证明了 ViGText 有能力抵御来自资源丰富且信息通达的对手的攻击。

TABLE IV 不同图像分辨率下以 SD 为测试集的性能(最高者以粗体显示)。

Resolution		450x450				512x512				550x550			
Resolution	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	
DCT	46.60	6.20	32.29	10.40	85.50	83.30	88.80	85.90	51.60	6.20	67.39	11.35	
DE-FAKE	93.40	93.90	92.97	93.43	92.40	91.10	94.00	92.50	93.45	94.10	92.89	93.49	
UnivCLIP	92.29	92.09	92.46	92.28	93.00	92.30	93.90	93.10	92.44	92.19	92.66	92.43	
ViGText	96.40	99.90	93.28	96.53	99.25	99.80	98.52	99.26	97.20	95.20	99.17	97.14	

继续进行稳健性评估,我们针对图像分辨率变化所引起的操控测试了 ViGText。这是稳健性的重要方面,因为实际应用中由于捕获条件多样性或后期处理步骤,输入的图像往往分辨率不同。表格 IV 展示了 SD 数据集的结果,该数据集的初始图像分辨率为 512x512。而表格 XI 则显示了 StyleCLIP 数据集的相应结果,其初始图像分辨率为1024x1024。

表 IV 的结果表明,在 SD 数据集的所有测试分辨率中, ViGText 的性能降级最小。这种一致的性能突显了其对分辨率变化的鲁棒性,准确率、精确率、召回率和 F1 分数均保持较高水平。值得注意的是,ViGText 在每个分辨率上都优于所有基线方法,这表明其基于图的框架有助于其有效适应图像分辨率的变化。

接下来,表格 V 和 VI 分别展示了在几何和外观变形操作下,所有评估方法在 SD 数据集上的表现。ViGText在这些变换中始终表现出优越的指标,表明其对空间畸变有强大的抗性。特别是,ViGText在遭受显著几何修改时,仍保持高精度、准确率、召回率和 F1 分数,展现了其在可能旋转或空间变换的现实场景中的适应性。为了完整性,

TABLE V 以 SD 作为测试集,不同几何变形操作的表现(最高的用粗体表示)。

Technique		Ro	tate		Scale and Translate				
recimique	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	
DCT	54.6	54.9	51.4	53.0	57.5	62.5	37.4	46.8	
DE-FAKE	86.8	81.7	95.0	87.8	89.2	86.0	93.6	89.6	
UnivCLIP	88.1	84.0	94.1	88.8	90.9	86.6	96.9	91.5	
ViGText	98.0	96.1	100.0	98.0	99.6	99.9	99.2	99.6	

我们还在附录 X 中包含了补充实验,涵盖了分辨率变化以及 StyleCLIP 数据集上的几何和外观变形操作,进一步展示了 ViGText 在不同数据集和操作类型中表现的一致性。

TABLE VI 针对不同外观变换操作,以 SD 作为测试集的表现(最高的用粗体标 出)。

Technique		Blur	ring		Brightness				
recinique	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	
DCT	80.00	67.00	90.54	77.01	81.2	78.7	85.6	81.9	
DE-FAKE	92.60	95.70	90.11	92.82	92.2	92.7	91.7	92.2	
UnivCLIP	92.74	96.39	89.84	93.01	91.4	90.9	92.2	91.5	
ViGText	97.60	99.90	95.42	97.66	99.6	99.2	99.9	99.6	

最后,表 VII 展示了在不同噪声水平下使用快速梯度符号法 (FGSM) [?] 和投影梯度下降 (PGD) [?] 攻击的对抗鲁棒性结果。在这里,ViGText 相较其他方法保持了明显的优势,在所有测试条件下都达到了最高的准确率。即使在较高的噪声幅度下,ViGText 也展示了相较竞争方法明显的改进空间。这种强大的对抗性韧性,加上对几何和外观转换的鲁棒性,确认了 ViGText 在实际处理中面对多样且具挑战性的操作时的适用性。

在这一小节中,我们调查 Q4: ViGText 对其开发过程中所做的设计选择有多敏感? 我们探讨了 ViGText 对图像分割为若干块的数量的敏感性,并评估了不同数据集和情景下块数量对性能的影响。

表格??总结了当我们改变块大小时,在SD和StyleCLIP数据集上的结果。结果表明,两个数据集上的性能在不同块大小下相对稳定,表明ViGText对此参数不太敏感。然而,当观察表格??中微调模型数据集的一般化性能时,我们发现当块大小减小时(也就是块数增加时)性能显著提高。

性能随着块数增加而提高,可以归因于这些图像中存在的伪影的性质。如图 8 所示,由 LoRA 微调模型 (图 8 (a))

TABLE VII 使用 FGSM 和 PGD 攻击的对抗图像准确率(最高的用粗体表示)。

Attack	Noise ( $\epsilon$ )	DCT	DE-FAKE	UnivCLIP	ViGText
Trouch	0.0001	82.41	88.21	82.37	96.43
FGSM	0.001	75.78	83.02	54.99	93.71
	0.01	71.09	71.48	37.19	89.19
	0.0001	35.16	63.52	61.84	91.46
PGD	0.001	16.24	61.04	56.55	87.83
	0.01	9.47	58.36	51.31	80.94
No Attack	_	85.50	92.45	93.04	99.25

和 FM 微调模型 (图 8 (b)) 生成的图像包含局部化的伪影,这些伪影在使用较小的块时更容易捕捉。当我们减小块大小时,图的空间粒度增加,这使得模型能够更好地定位和表示这些失真。较小的块也创造了更多的图节点,这提高了模型对局部区域细微变化的敏感性,进一步增强了其一般化能力。我们注意到不同的行为,如表?? 所示,在StyleCLIP 数据集的对抗性操控测试集合上的表现呈现出相反的趋势,较大的补丁尺寸产生更好的结果。这种行为可能是因为对抗性图像(如图 8 (c)) 缺乏明显的伪影,而较小的补丁可能无法捕捉到有意义的特征。对于这些图像,单个补丁内的信号较弱,图节点数量的增加可以稀释信号,这使得 ViGText 更难区分真实和伪造内容。此外,较小的补丁降低了图对整体模式的捕捉能力,而整体模式对于检测旨在规避的对抗性操控至关重要。

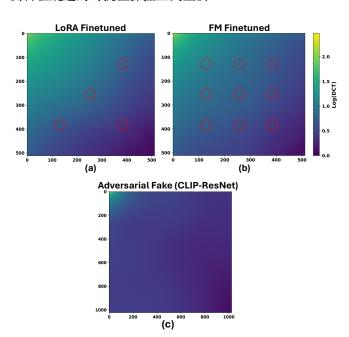


Fig. 8. 对数 DCT 频谱显示: (a) 使用 LoRA 微调模型生成的图像中的伪影, (b) 使用 FM 微调模型生成的图像中的伪影, (c) 对抗生成图像中没有伪影。红色圆圈展示了这些伪影。

ViGText 对补丁大小的敏感性与所分析图像的性质密切相关。对于充满伪影的图像,较小的补丁通过捕获局部失真来提高性能,而对于无伪影的对抗性图像,较大的补丁更为有效,因为它们保留全局上下文。这表明一种自适应补丁策略,即根据输入图像的特征动态调整补丁大小,可以进一步增强 ViGText 的鲁棒性和泛化能力。

#### E. 成本分析

最后,我们回答 Q5: 运行 ViGText 的成本如何? 我们将 ViGText 与 UnivCLIP 进行比较, UnivCLIP 在先前的实验中展示了第二好的整体性能。时间成本被测量为预处理图像、运行模型推理所需的总时间,并且对于 ViGText,还包括图的构建。这些结果是对 2000 个测试样本的平均值。

平均而言, UnivCLIP 从预处理到推理每张图像需用时 1.650 秒, 而 ViGText 仅稍多, 用时 1.755 秒。这 0.105 秒的时间成本微增, 突显了 ViGText 的效率。结果表明, ViGText 在具备卓越检测性能的同时仅需增加很少的计算

成本。先进 VLLMs 的使用,如 Qwen2 [?] ,大大提升了这种效率。Qwen2-VL-7B-Instruct 是一个紧凑而强大的模型,能够对图像内容进行深入理解,并以低计算开销生成详细的文本说明。通过这些进步,ViGText 能够将高质量的解释融入其基于图的框架中,同时保持具有竞争力的运行时间。

这些发现表明, ViGText 不仅在检测准确性和鲁棒性方面表现出色, 而且在实际部署中仍然实用, 因为它以几乎可以忽略不计的时间成本增加实现了最先进的性能。

## V. 相关工作

Deepfake 检测。最近的 deepfake 检测方法可以根据模 型架构大致分为两种主要方法: 基于 CNN 的方法和基于 传统机器学习模型的方法。基于 CNN 的方法侧重于利用 深度学习架构来检测假图像。例如,[?] 探索了假图像的特 征,这些特征可以实现跨不同生成模型和数据集的检测。 这项研究引入了一种基于局部接收域的分类器,强调局部 图像伪影而不是整体结构。[?] 提出了 Gram-Net, 这是一 种通过集中于全局纹理特征来增强假脸检测的模型,提高 了对各种 GAN 模型和图像失真效果的鲁棒性和泛化能力。 [?] 同样旨在提高鲁棒性和泛化能力,但采用不同的方法, 通过超级分辨率、去噪和上色等任务对图像进行重合成, 而不仅仅依赖于频率伪影。此外,[?]针对面部视频伪造, 使用两种轻量级的 CNN 架构来分析介于细节与高层内容 之间的中观特性。传统的基于机器学习的方法最近将重点 从复杂的深度学习架构转向高级特征提取技术,以训练更 简单的模型进行深度伪造检测。在[?]中,作者讨论了使 用频域特征来检测扩散模型的深度伪造, 并强调这些特征 如何能够揭示在空间域中不易看出的微妙伪影。同样, [?] 提出将文本提示特征和生成图像特征组合作为简单分类器 的输入,使其能够学习文本和视觉内容之间的关联,以便 准确检测假图像。最后,[?]介绍了一种通过利用一个并非 专门用于深度伪造检测的特征空间来检测由各种模型生成 的假图像的方法,其中包括那些在训练期间未见过的模型。 [?] 中的作者研究了前面提到的方法,并强调了它们在面 对新威胁时的脆弱性。这些威胁主要涉及易于获取的用户 定制生成模型和使用基础模型创建的对抗性深度伪造。研 究表明,当面对由用户定制变体生成的图像时,这些方法 难以泛化。此外,试图检测使用大型基础模型生成的对抗 性图像时性能明显下降。

### VI. 讨论

研究的影响。本研究的影响在于其对深度伪造检测技术的重大进步。通过采用基于图形的框架整合视觉和文本数据,这一方法直接提高了检测方法在应对不断发展的生成模型带来的挑战时的可靠性和稳健性。该方法提高了对用户定制模型的泛化能力,并增强了对复杂对抗攻击的抵御能力。这些进步对于保护媒体的真实性和可信性尤为关键,因为媒体正面临来自 AI 生成的合成内容日益增长的威胁。

除了深度伪造检测之外,该研究的方法可以适用于需要辨别真伪或辨别有害与有益现象的应用。与文本解释一起处理数据,通过提供背景洞察和增强可解释性丰富了分析。例如,在毒性化学物质的识别中,解释可以详细说明与毒性相关的特征。在药物发现中,将生物图像与文本描述结合,可以揭示结构之间的关系。同样,在内容审核、文件

验证和假新闻检测中,文本解释可以补充视觉数据,从而 实现更全面和可靠的评估。

局限性。虽然这项研究取得了显著进展,但必须承认几个局限性。首先,当前的威胁模型和评估集中于完全合成的图像,包括潜在空间操作,如由 StyleCLIP [?] 生成的图像。尽管这些可以模拟局部语义编辑,但它们并不代表现实世界中的部分操控,如换脸、重现或拼接内容。其次,该框架目前仅限于视觉模式,尚不支持音频或视频输入。扩展到这些模式需要解决新的挑战,例如建模时间动态、对展到这些模式需要解决新的挑战,例如建模时间动态、对来异步多模态信号以及处理录制质量和噪声的可变性。最后,从技术角度来看,图构建中的固定大小补丁使用和词到补丁链接策略可能会限制对图像内容复杂性的适应性,并降低上下文精度,指出了未来改进的明确方向。

未来工作。未来的方向包括通过考虑上下文依赖性更有 效的自适应链接机制来增强图像块与文本解释之间的对 齐。自适应拼接策略,其中块大小和图连通性根据图像内 容动态调整,可能改善可扩展性和定位精度。引入异构图 神经网络是另一个有前途的步骤,通过建模多样化的节点 和边类型来实现更具表现力的表示。虽然这项工作集中于 图像,但将该方法扩展到其他模态如音频和视频则带来了 相当大的挑战。基于视频的深度造假需要时间建模,以跟 踪跨帧的空间和运动一致性,而基于音频的检测则涉及提 取与视觉或文本线索一致的重要声学特征。此外,多模态 融合涉及同步性、信号质量变化以及模态特异性对抗性攻 击的问题。解决这些复杂性需要重新设计模型架构,以处 理时间序列数据,支持多流对齐,并在跨模态不一致性方 面保持稳健。最后,基于频域稳健性,未来工作可以探讨 使用学习的变换基或多分辨率表示, 以更好地泛化合成和 真实世界的操纵,并提高对抗扰动的抵抗力。

在这项工作中,我们介绍了一种名为 ViGText 的框架, 这种框架用于深度伪造检测,结合了来自 VLLMs 的解释 和视觉数据,以双图结构整合在一起。这种新颖的方法解 决了现有方法中的关键局限性, 因为它在微调和用户定 制的深度伪造上表现出卓越的泛化能力,对抗干扰有很强 的抵抗力, 并适应多样的测试场景。通过结合空间和频率 域特征与详细文本解释的基于图形的表示, ViGText 实现 了最先进的检测性能。ViGText 的独特性在于无缝地连接 视觉和文本模态, 使其能够检测传统方法难以处理的微妙 伪影和不一致性。尽管 ViGText 具备先进能力,它仍然 保持计算效率,确保在实际中的可行性。这使得 ViGText 成为一个可扩展的解决方案, 能够应对生成模型和对抗 性技术快速发展带来的挑战。随着合成媒体技术的不断发 展,对强大和适应性强的检测系统的需求变得越来越迫切。 ViGText 不仅解决了深度伪造检测日益增长的挑战,还为 该领域的未来进步奠定了坚实的基础。它在保护数字世界 免受欺骗和确保在线内容的信任方面发挥了重要作用。

这项研究不使用任何涉及隐私和安全的敏感数据。我们使用开源数据集进行实验。该研究是在承诺维护最高道德标准的情况下进行的。所采用的方法经过精心设计,以确保研究尊重个人隐私和权利,避免偏见,并将潜在损害降至最低。研究中使用的数据是负责任地获取和管理的,遵循以机密性和信息的道德使用为优先的道德准则。研究旨在为朝向安全、可靠和可信的人工智能进行的反深伪研究作出积极贡献。

是一个图 (如图 ?? 所示),展示了在将图像提供给视觉 大语言模型以生成相应解释时使用的结构化模板。

### VII. 稳定扩散 3.5 LoRA 模型

表 VIII 提供了从 Hugging Face [?] 获取的 Stable Diffusion 3.5 的 LoRA 微调模型的概述,这些模型用于生成本研究中的泛化评估扩展。表格中包含这些模型的直接链接,许多模型被广泛采用并拥有数千次下载。

TABLE VIII 本工作中使用的 STABLE DIFFUSION 3.5 LORA 模型。

Model Description	Link
Ancient Stlye	https://huggingface.co/reverentelusarca/ancient-style-sd35
Anime	https://huggingface.co/prithivMLmods/SD3.5-Large-Anime-LoRA
Chinese Line Art	https://huggingface.co/Shakker-Labs/SD3.5-LoRA-Chinese-Line-Art
Futuristic Bronze Colored	https://huggingface.co/Shakker-Labs/SD3.5-LoRA-Futuristic-Bzonze-Colored
Photorealistic	https://huggingface.co/prithivMLmods/SD3.5-Large-Photorealistic-LoRA
Pixel Art	https://huggingface.co/nerijs/pixel-art-3.5L
Red Light	https://huggingface.co/Shakker-Labs/SD3.5-LoRA-Linear-Red-Light
Rustic Whimsy	https://huggingface.co/crystalwizard/Rustic-Whimsy-SD3.5-Large-Lora

# VIII. 补充结果: 最先进的图像生成 API 和攻击

TABLE IX VIGTEXT 在由最先进的扩散 API 生成的图像上的表现

Image Generation API	Accuracy	Precision	Recall	F1
OpenAI	99.49	99.94	98.96	99.48
Google Gemini	96.98	98.98	95.12	97.01

表格 IX 报告了 ViGText 在由两个最先进的商业扩散图像生成 API 合成的图像上的表现: OpenAI 的 Image-1 和Google 的 Gemini 2.5。这些实验直接解决了关于 ViGText 在经常用于创建高保真深度伪造的最新生成模型下的有效性问题。ViGText 实现了接近完美的检测率,在 OpenAI 生成的图像上保持 99.49 % 的准确率,在 Gemini 图像上保持 96.98 % 的准确率。

TABLE X 在 CHIMERA 重捕获 + DEEPFAKE 攻击下 VIGTEXT 的表现

Tested Images	Accuracy	Precision	Recall	F1
Benign Images	99.49	99.96	99.12	99.56
Attacked Images	82.08	83.00	81.51	82.25

表 X 展示了 ViGText 在更复杂的对抗场景中的韧性,特别是由 Park 等人介绍的双阶段 "Chimera" 攻击。该攻击结合了物理重捕和用于欺骗图像真实性检测器的深度伪造操控。尽管存在这些挑战性的条件, ViGText 在被攻击图像上的准确率仍能保持 82.08 %,显著高于 Chimera 研究中评估的领先公共检测器报告的 58.5–69.0 % 范围。这确认了 ViGText 对抗不断出现的混合深度伪造威胁的适用性。

## IX. 补充结果: 样本图像

图 9 显示了来自 SD 数据集的样本图像,这些样本图像由 ViGText 唯一正确分类,而所有其他基线方法都将它们分类错误。图 9 的第一行显示了伪样本,第二行显示了真实样本,突出显示了 ViGText 在区分真实内容和生成内容方面的卓越检测能力。



Fig. 9. 从 SD 数据集中提取的样本图像, 仅由 ViCText 正确分类, 而 所有其他基线均未能成功。第一行显示伪造样本, 第二行显示真实样本。

TABLE XI 在不同图像分辨率下使用 STYLECLIP 作为测试集的性能(最高的以粗 体显示)。

Resolution	900x900				1024x1024				1100x1100			
Resolution	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
DCT	61.20	95.40	56.65	71.08	98.80	98.20	99.40	98.80	68.40	46.80	82.39	59.69
DE-FAKE	74.25	71.60	75.61	73.55	74.00	75.30	71.50	73.30	74.20	71.60	75.53	73.51
UnivCLIP	93.09	92.49	93.62	93.05	93.00	93.80	92.10	92.90	93.09	92.09	93.97	93.03
ViGText	99.00	99.90	98.04	99.01	99.60	99.90	99.21	99.60	99.20	99.99	98.43	99.21

# X. 补充结果: STYLECLIP 数据集上的分辨率和扭曲操作

类似于表格 IV , 表格 XI 显示在测试的分辨率内 ViGText 在 StyleCLIP 数据集上保持了高检测性能,包括原始分辨率 1024x1024 和调整后的分辨率 900x900 和1100x1100。即使在非原生分辨率下, ViGText 在所有方法中也达到了最高的性能,各种指标如准确率和 F1 分数在所有情况下均超过 99 %。尽管一些基线在非原始分辨率下表现出明显的退化, ViGText 展现了显著的鲁棒性,这表明其对不同输入尺寸的适应性。

TABLE XII 在不同几何变形操作中的测试集性能使用 STYLECLIP(最高的用粗体 标出)。

Technique		Rot	ate		Scale and Translate				
	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	
DCT	52.2	53.8	31.2	39.5	70.4	79.8	54.6	64.8	
DE-FAKE	68.8	75.5	55.7	64.1	69.4	71.5	64.5	67.8	
UnivCLIP	93.0	93.8	92.2	92.9	87.0	80.2	98.4	88.4	
ViGText	94.8	94.4	95.2	94.8	99.6	87.9	98.4	92.4	

表格 XII 和 XIII 分别展示了所有评估技术在 StyleCLIP 数据集下几何和外观畸变操作中的性能。结果进一步突出了 ViGText 在不同类型失真中的多功能性和稳健性。

在表 XII 中,评估了旋转和缩放-平移操作, ViGText 在 所有类别中都取得了最高的指标,在旋转下保持了 94.8 % 的准确率,在缩放和平移下则达到了令人印象深刻的 99.6 %。这些结果显示了 ViGText 对大幅几何变换的广泛适应能力,考虑到图像处理中无论是良性还是对抗性的空间操作都非常普遍,这是一项至关重要的能力。

TABLE XIII 以 STYLECLIP 为测试集在不同的基于外观的扭曲操作中的表现(最高的用黑体标出)。

Technique	Blurring				Brightness			
	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
DCT	65.80	99.20	59.47	74.36	94.1	89.6	99.8	94.4
DE-FAKE	73.90	72.10	74.79	73.42	73.2	76.4	67.3	71.5
UnivCLIP	93.19	92.89	93.46	93.17	89.7	89.3	90.2	89.8
ViGText	99.4	99.8	99.01	99.4	99.6	99.2	99.9	99.6

与此同时,表格 XIII 检查了对基于外观的更改的弹性,例如模糊和亮度调整。在此,ViGText 实现了近乎完美的性能,持续以较大优势超越其他方法。例如,在模糊情况下记录了 99.4 % 的准确率,在亮度变化下记录了 99.6 % 的准确率,同时具有平衡的精确度、召回率和 F1 分数。这些结果表明,ViGText 不仅在几何变形下表现出色,而且在各种光度扰动下保持高保真度,增强了其在图像质量和照明经常波动的现实场景中的实用价值。