XXXX

基于跨模态语义对齐的电信诈骗图文一致性检测方法

摘　要：随着生成式AI发展，电信诈骗日益呈现视觉化与复合化趋势，诈骗分子通过伪造转账截图、公文等图像搭配误导文本，构成“图文并悖”欺诈，传统检测系统面临挑战。为此，本文提出一种基于深度跨模态语义对齐的图文一致性检测方法：先利用微调的YOLOv8定位图像中关键视觉元素，并解析文本实体与欺诈意图；再借助预训练CLIP模型将视觉与文本信息编码至同一语义空间；最后通过计算特征向量余弦相似度并结合动态阈值评估语义矛盾。在自建电信诈骗图文数据集上的实验表明，该方法F1值达92.7%，较特征拼接基线提升19.6%，且对噪声、裁剪等干扰具有良好鲁棒性，为自动化、高精度的混合型诈骗识别提供了可解释解决方案。

关键词：电信诈骗检测；跨模态对齐；语义一致性；CLIP；目标检测

中图分类号：TP393

文献标志码：A

doi: 10.11959/j.issn.1000−0801.

Cross-Modal Semantic Alignment-Based Method for Detecting Text-Image Consistency in Telecom Fraud

Abstract: With the advancement of generative AI, telecom fraud is increasingly exhibiting visual and composite trends. Fraudsters combine forged images, such as transfer screenshots and official documents, with misleading text to form "text-image contradiction" scams, posing significant challenges to traditional detection systems. To address this issue, this paper proposes a deep cross-modal semantic alignment-based method for detecting text-image consistency. The method first employs a fine-tuned YOLOv8 model to localize and extract key visual elements, such as logos and seals, from images, while simultaneously parsing key entities and fraudulent intents from the text. Subsequently, a pre-trained CLIP model is utilized to encode the extracted visual regions and text information into the same semantic space, obtaining comparable feature vectors. Finally, the semantic contradiction between text and images is quantitatively assessed by computing the cosine similarity of the visual and textual feature vectors combined with an adaptive dynamic threshold. Experimental results on a self-built dataset of telecom fraud text-image pairs demonstrate that the proposed method achieves an F1-score of 92.7%, representing a 19.6% improvement over the feature concatenation baseline. Furthermore, it exhibits good robustness against common adversarial interferences such as noise addition and image cropping. This study provides an efficient and interpretable solution for automated, high-precision detection of hybrid telecom fraud involving both text and images.

Key words: Telecom fraud detection, Cross-modal alignment, Semantic consistency, CLIP, Object detection

收稿日期：XXXX−XX−XX；修回日期：XXXX−XX−XX

1 引言

近年来，生成式人工智能技术的突破性进展在推动数字内容创新的同时，也被不法分子恶意利用，致使电信诈骗活动的手段与形式发生深刻演变。诈骗模式已从传统单一的话术诱导演进为一种高度视觉化与复合化的新型欺诈[1]。犯罪分子能够便捷地利用深度伪造与图像合成技术，批量制作出以假乱真的政务公文、银行转账截图、钓鱼网站界面等视觉素材。这些高度仿真的伪造图像与精心编排的、具有强误导性的文本描述相结合，构成了“图文并悖”的欺诈组合，极大地增强了诈骗活动的迷惑性和欺骗性，对社会经济安全构成严重威胁[2,3]。

面对这一新型挑战，当前主流的电信诈骗检测系统在应对上显得力不从心。依赖单一信息模态的检测手段，例如基于文本关键词过滤或基于图像篡改痕迹识别[4,5]的方法，由于无法建立并分析图文之间的语义关联，对这类复合欺诈的检测效能急剧下降[6]。部分研究尝试融合图文信息，但多采用特征拼接或决策级融合等浅层策略，未能实现跨模态信息的深度语义对齐，因而难以精准识别图文之间深层次的逻辑矛盾[7,8,9]。其根本症结在于，现有系统普遍缺乏对图像内容与伴随文本之间内在语义一致性进行联合分析与判断的能力。

与此同时，在多模态人工智能研究领域，基于对比学习的跨模态预训练模型（如CLIP）通过在海量互联网图文数据上进行训练，展现了出色的图像与文本语义对齐能力[10]。然而，直接将此类通用模型应用于电信诈骗检测这一特定而细粒度的任务时，仍面临显著挑战。一方面，对整幅图像进行全局编码会稀释其中小尺寸、高语义密度的关键视觉元素（如标识、印章）的信号，而这些元素往往是识别欺诈的关键线索[11]；另一方面，通用模型未针对诈骗场景中特有的视觉对象（如伪造UI组件）和文本诱导模式进行优化，其领域适应性有待提升[12]。

综上所述，现有方法在应对“图文并悖”电信诈骗时，存在语义对齐不足、领域适应性弱、可解释性差等局限。为填补这一研究与实践空白，本文提出一种基于跨模态语义对齐的图文一致性检测方法。本研究工作的核心贡献在于：

（1）提出一个面向诈骗场景的“关键区域检测—深度语义对齐”框架。该框架通过微调YOLOv8模型定位并提取图像中的关键视觉元素，进而利用CLIP模型将视觉信息与文本信息编码至同一高层语义空间，实现对图文语义矛盾的精准度量。

（2）构建并公开了一个覆盖多种伪造场景的电信诈骗图文数据集（FraudBench），并对图文对进行了“语义一致”与“语义矛盾”的精细标注，为后续相关研究提供了基准数据支持。

（3）通过系统的实验验证，从准确性、鲁棒性、可解释性等多个维度证明了所提方法的优越性，并提供了详尽的消融实验分析，验证了各核心组件的有效性。

2 相关工作

针对电信诈骗的检测技术，现有研究主要围绕单一模态分析和多模态信息融合两条路径展开。随着诈骗手段向多模态复合形态演进，单纯依赖文本或图像的方法已显现出明显不足，而多模态融合方法在深层语义对齐与领域适应性方面仍面临挑战。

2.1 单一模态诈骗检测

传统检测系统大多依赖于对单一模态信息的分析。文本分析方法主要借助自然语言处理技术，通过对通信内容进行风险分类、线索词提取或情感分析来识别欺诈意图[13]。这类方法虽能捕捉文本层面的异常模式，但完全忽视了伴随图像中可能存在的伪造视觉证据。另一方面，图像真

伪检测技术则专注于识别图像自身的篡改痕迹，例如针对深度伪造人脸的生物特征分析[14]，或检测图像拼接、复制-移动等操作留下的底层特征不一致性[15]。然而，这类方法缺乏对图像语义内容的理解，更无法判断其与上下文文本描述是否逻辑自洽。因此，无论是仅分析文本还是仅检验图像，单模态方法对“图文结合”的复合欺诈手段均难以有效应对[16]。

2.2　多模态信息融合方法

为应对多模态欺诈，早期研究尝试融合图文两种信息。主流做法可归类为特征级拼接或决策级融合[17]。例如，有工作将卷积神经网络（CNN）提取的图像全局特征与BERT模型提取的文本特征向量进行拼接，随后输入到一个分类器中进行联合欺诈识别[18]。然而，由于图像和文本特征最初来自相互独立的语义空间，这种简单的拼接操作并未实现真正的特征对齐，模型需要依赖大量的标注数据来隐式学习模态间的关系，且其决策过程通常缺乏可解释性。近年来，基于Transformer架构的多模态预训练模型，如ViLBERT[19]和UNITER[20]，通过跨模态注意力机制实现了图文特征的深层交互，性能有所提升。但这类模型往往流程复杂，计算开销较大，在需要高并发、实时响应的诈骗检测实际场景中部署存在困难[21]。此外，尽管基于知识图谱的方法能从关联分析角度辅助识别诈骗网络，或AIGC检测技术可用于鉴别伪造内容，但这些工作大多侧重于单一模态或宏观关联分析，未能直接解决图文混合欺诈中的语义一致性判定问题。

2.3　基于对比学习的跨模态模型

CLIP模型的提出为跨模态理解提供了新范式。该模型通过在超大规模的互联网图文对上进行对比学习，将图像和文本编码到统一的语义空间中，从而实现了强大的零样本迁移能力。这一特性使其在诸多需要图文匹配的任务中表现出色[22]。然而，直接将原始CLIP应用于诈骗图文一致性检测这一细粒度任务仍存在局限。其一，对整图进行编码会稀释小尺寸关键视觉元素（如Logo、印章）的语义信号，而这些元素在诈骗判定中至关重要；其二，CLIP作为通用模型，并未针对诈骗场景中特有的视觉模式进行优化，其识别相关视觉概念的精度有待提升。近期，也有研究尝试将对比学习思想与小样本学习结合用于诈骗文本分类[23]，或利用大语言模型分析诈骗事件风险[24]，但这些工作尚未深入解决跨模态的语义对齐与矛盾判定问题。与此同时，多模态大语言模型（MLLM，如LLaVA、Qwen-VL）的兴起为图文理解提供了新的端到端范式，但其巨大的参数量和推理成本，以及在特定领域（如小目标伪造区域识别）的细粒度对齐能力，仍是在线实时检测场景中需要权衡的问题。

2.4　诈骗检测中的辅助技术研究

除直接的图文内容分析外，网络安全领域还有一系列辅助性研究为反诈体系提供支撑。例如，基于威胁环境感知的动态检测框架旨在从更宏观的层面评估风险；针对涉诈网站的检测与分类技术则专注于URL和页面结构特征；知识图谱被用于分析和补全诈骗网络中的关联链接；此外，针对AIGC生成伪造内容的检测与防御技术也获得了广泛综述与探讨。这些工作从不同维度丰富了反诈技术体系，但大多侧重于单一技术路径或特定数据类型，未能系统性地解决图文混合欺诈中最为核心的语义一致性判定问题。

综上所述，现有方法在应对“图文并悖”型电信诈骗时，或在模态覆盖上存在缺失，或在融合深度上有所不足，亦或缺乏对特定场景的适应能力。本文工作的核心创新在于，将面向诈骗场景的细粒度关键视觉信息检测，与CLIP模型强大的深度语义对齐能力相结合，以此实现对图文语义矛盾高效、精准且可解释的检测。

3 系统设计与实现

本章节详细阐述本文提出的基于跨模态语义对齐的图文一致性检测方法。如图1所示，整体架构采用端到端的管道设计，由四个核心模块依次构成：（1）多模态数据预处理与对齐模块；（2）面向诈骗场景的关键区域检测与文本语义解析模块；（3）基于对比学习的深度跨模态语义编码与融合模块；（4）可解释的图文一致性度量与自适应决策模块。本方法的核心创新在于，将面向诈骗场景的细粒度视觉元素检测，与基于大规模预训练模型的深度语义空间对齐相结合，从而实现对图像中非生物关键对象（如Logo、印章、界面组件）与诱导性文本之间深层语义矛盾的精准、高效且可解释的度量。

3.1 整体架构

系统整体工作流程如图1所示，主要包含四个阶段：1）多模态数据采集与预处理；2）关键视觉区域检测与文本结构化特征提取；3）跨模态深度语义编码与特征对齐；4）图文不一致性量化与风险判定。

给定一个待检测的图文对(I, T)，其中I为输入图像（例如伪造的银行转账截图或钓鱼网站界面），T为与之关联的文本内容（例如诱导性描述或聊天记录）。系统的目标是输出一个量化的风险评分 $R \in [0, 1]$，并提供可解释的证据集合 $\mathcal{E}$，用以清晰表明图像I与文本T之间存在的语义不一致性。该过程可形式化定义为：

$$Output = \mathcal{F}(I, T; \Theta) = (R, \mathcal{E}) \quad (1)$$

其中，$\mathcal{F}$ 为本方法定义的函数，Θ 为模型参数，$\mathcal{E}$ 为可解释的证据集合（如：“检测到‘银行Logo’视觉元素，与文本中‘虚拟货币投资’的语义存在矛盾”）。

系统采用级联式设计，各模块依次连接，前一模块的输出作为后一模块的输入。这种设计不仅保证了信息处理流程的连贯性与高效性，也通过模块化解耦增强了系统的可扩展性和可维护性。

3.2 领域自适应的细粒度视觉关键区域检测

在电信诈骗的图文素材中，具有高度判别性的视觉信息往往集中于尺寸较小但语义密集的区域。如下图2所示，例如文件角落的公章、界面一角的银行Logo或安全认证图标。若直接对整幅图像进行全局编码，这些关键信号的语义容易被背景信息稀释。为此，本文设计了一种两阶段的关键视觉区域检测与筛选策略。

第一阶段：基于微调YOLOv8的细粒度目标检测。我们使用自建的诈骗图文数据集 D_{fraud} 对YOLOv8模型进行监督微调。该数据集包含了针对诈骗场景定义的视觉类别

C_{visual} =

Logo, Seal, Certificate, Face, Amount_Area, UI_Button, Security_Badge

的边界框标注。为提高模型对诈骗场景中小目标

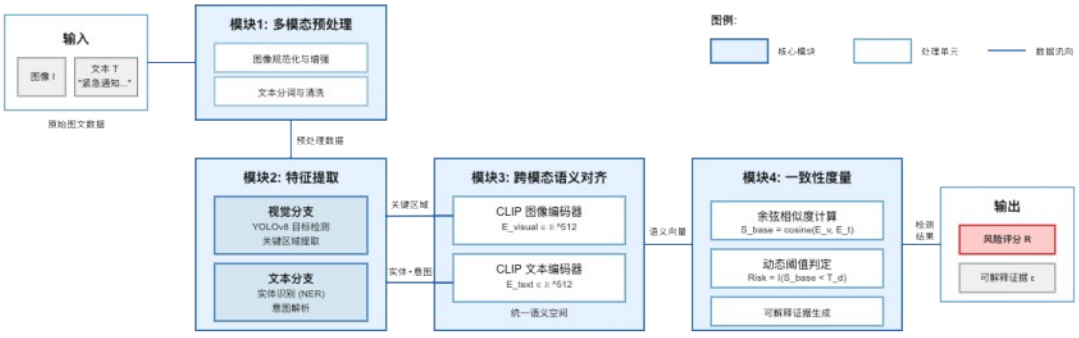

图1 基于跨模态语义对齐的电信诈骗图文一致性检测系统架构

的检测能力，微调过程中采用了两种优化策略：一是数据增强阶段引入针对小目标的随机缩放与局部裁剪；二是在损失函数中为小目标类别（如 Seal）分配更高的权重 λ_{small}，以缓解类别不平衡问题。检测器最终输出一组候选区域 $B=\{b_i \mid i=1,...,N\}$ 及其对应的类别标签 c_i 与置信度 s_i。

YOLOv8 微调总损失函数：

$$\mathcal{L}_{YOLO}=\lambda_{cls}\mathcal{L}_{cls}+\lambda_{box}\mathcal{L}_{box}+\lambda_{df}\mathcal{L}_{df}+\lambda_{small}\mathcal{L}_{small} \quad (2)$$

第二阶段：基于语义重要性的区域筛选。并非所有检测到的区域都具有同等的欺诈指示意义。为此，我们设计了一个轻量级的区域重要性评分模块。对于每个候选区域 b_i，其重要性分数 γ_i 由检测置信度 s_i 和一个基于先验知识的类别重要性函数 $\mathcal{S}_{prior}(c_i)$ 共同决定：

$$\gamma_i=\alpha\cdot s_i+\beta\cdot\mathcal{S}_{prior}(c_i) \quad (3)$$

其中，s_i 是检测置信度。$\mathcal{S}_{prior}(c_i)$ 是一个基于先验知识的类别重要性函数，我们根据诈骗分析经验预先设定（例如，Seal 和 Bank_Logo 的权重高于普通 UI_Button）。α 和 β 是可调的超参数。最终，我们选取 γ_i 排名前 K 的区域构成关键视觉区域集合 $V=\{v_1,v_2,...,v_K\}$。

该模块的创新性在于其“领域适应性”与“语义导向的筛选机制”。通过对 YOLOv8 进行场景特定的微调，使其精准捕捉诈骗图文中的典型视觉模式；再通过重要性评分，模拟人工审核中对高欺诈风险证据的优先关注，从而显著提升了信息提取的针对性和效率。

3.3　文本语义解析与关键信息抽取

诈骗文本通常包含特定的诱导模式（如制造紧迫感、假冒权威）和关键实体信息（如机构名、金额）。为与视觉侧的关键信息对齐，我们设计了并行的文本解析流程，同时抽取表层实体与深层意图。实体识别与意图分类的联合损失函数：

$$\mathcal{L}_{text}=\mathcal{L}_{ner}+\lambda_{intent}\mathcal{L}_{intent} \quad (5)$$

其中，$\mathcal{L}_{ner}$ 为命名实体识别的交叉熵损失，$\mathcal{L}_{intent}$ 为意图分类的损失，λ_{intent} 为平衡系数。

实体级信息抽取：采用预训练的 BERT 模型结合序列标注架构，进行命名实体识别（NER），抽取以下类别的实体：ORG（机构）、MONEY（金额）、DATE（日期）、PRODUCT（产品/服务）。这些实体构成了文本的“事实骨架”。

意图级语义理解：为捕捉文本的欺诈性意图，我们构建了一个诈骗意图词典 L_{fraud}，包含“高收益”、“官方认证”、“唯一名额”等数百个关键短语。通过模式匹配与基于 BERT 的句子意

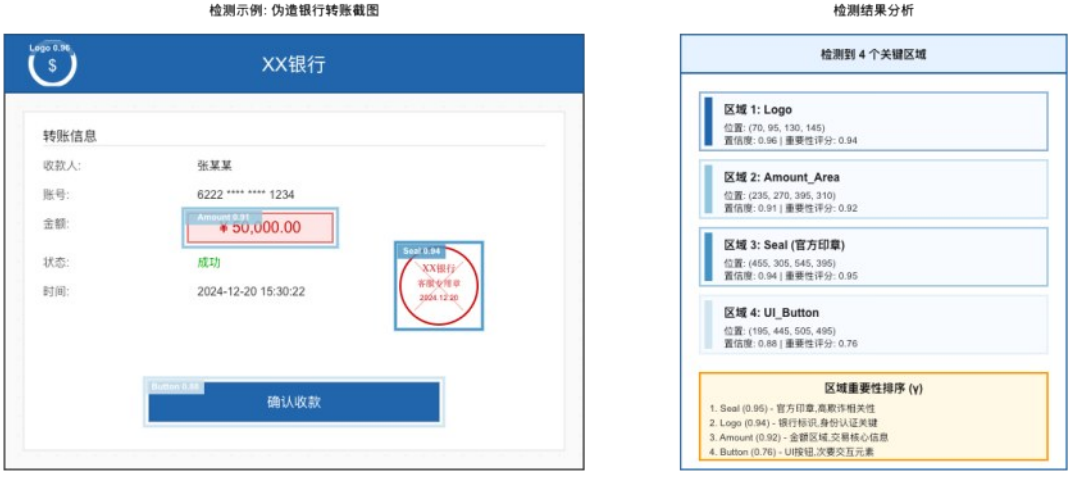

图2　基于 YOLOv8 的诈骗场景关键区域检测示意

图分类相结合，生成一个意图强度分数 ρ_{intent} 及对应的意图关键词集合 K_{intent}。

最终，文本侧的特征由实体集合 E_{entity} 和意图关键词集合 K_{intent} 共同表示。这种“实体-意图”双流解析方式，显式地分离和强化了文本中与欺诈风险评估最相关的成分，为后续与视觉元素的精准语义对齐奠定了基础，并有效过滤了无关文本信息的干扰。

3.4 基于对比学习的跨模态语义对齐与融合

本模块是整个方法的核心，目标是将视觉区域集合 V 和结构化文本信息 (E_{entity}, K_{intent}) 映射到一个统一的、可度量的语义空间。

视觉语义编码：对于每个关键视觉区域 v_k，我们使用 CLIP 的图像编码器 $Enc_I(\cdot)$ 提取其深度特征向量 $f_k^v \in \mathbb{R}^{512}$。为聚合 K 个区域的全局视觉语义，本文提出一种注意力加权的池化策略，而非简单的平均池化：

$$a_k = \frac{\exp(w^T f_k^v)}{\sum_{j=1}^{K} \exp(w^T f_j^v)}, E_{visual} = \sum_{k=1}^{K} a_k \cdot f_k^v \quad (6)$$

其中，w 是一个可学习的权重向量。该机制使模型能够自适应地关注与诈骗语义更相关的视觉区域。文本特征编码与提示构造：

$$T_{prompt} = Concat(E_{entity}, K_{intent}), \quad E_{text} = Enc_T(T_{prompt}) \quad (7)$$

$Concat$ 表示将实体集合 E_{entity} 和意图关键词集合 K_{intent} 组合成自然语言提示，Enc_T 为 CLIP 文本编码器。

文本语义编码：为充分利用 CLIP 的文本理解能力，我们将结构化的实体和意图信息合成为一段连贯的自然语言描述，例如：“文本提及[工商银行]，承诺[10 万元]收益，并包含‘高回报、稳赚不赔’等表述”。将该描述句输入 CLIP 的文本编码器 $Enc_T(\cdot)$，得到文本特征向量 $E_{text} \in \mathbb{R}^{512}$，工作流程如下图 3 所示。

语义空间对齐的理论基础：CLIP 模型通过在数亿图文对 (I_i, T_i) 上优化对比损失进行预训练：

$$\mathcal{L}_{CLIP} = -\frac{1}{N}\sum_{i=1}^{N}\left[\log \frac{\exp(\text{sim}(I_i, T_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(I_i, T_j)/\tau)} + \log \frac{\exp(\text{sim}(I_i, T_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(I_j, T_i)/\tau)}\right] \quad (8)$$

其中 $\text{sim}(\cdot,\cdot)$ 为余弦相似度，τ 为温度参数。此过程迫使模型将语义相关的图文对在特征空间中的距离拉近，不相关的推远。因此，Enc_I 和 Enc_T 的输出空间天生对齐，为本任务中的一致性度量提供了理想的基础。

CLIP 模型通过在海量互联网图文对上优化对比损失进行预训练，其图像编码器和文本编码器的输出空间天然对齐，语义相关的图文对在特征空间中距离更近。这为本任务中的一致性度量提供了理想的基础。

本模块的创新在于“结构化文本提示构建”与“注意力视觉聚合”。前者将离散的文本信息转化为 CLIP 擅长的自然语言句子；后者实现了对多区域视觉信息的智能融合。两者结合，确保了跨模态特征在高质量语义层面进行有效比对。

3.5 可解释的图文不一致性度量与自适应决策

在共享语义空间中，计算全局视觉向量 E_{visual} 与文本向量 E_{text} 之间的余弦相似度作为基础一致性分数：

$$S_{base} = \frac{E_{visual} \cdot E_{text}}{\| E_{visual} \| \ \| E_{text} \|} \quad (9)$$

细粒度矛盾诊断：为提供可解释的证据，我们不仅计算全局相似度，还计算每个关键视觉区域 v_k 与文本的相似度 $S_k = \text{cosine}(f_k^v, E_{text})$。设定一个矛盾阈值 η，若 $S_k < \eta$，则标记区域 v_k 为矛盾区域。系统最终输出矛盾区域列表及其类别标签 c_k，作为风险证据 $\mathcal{E}$。

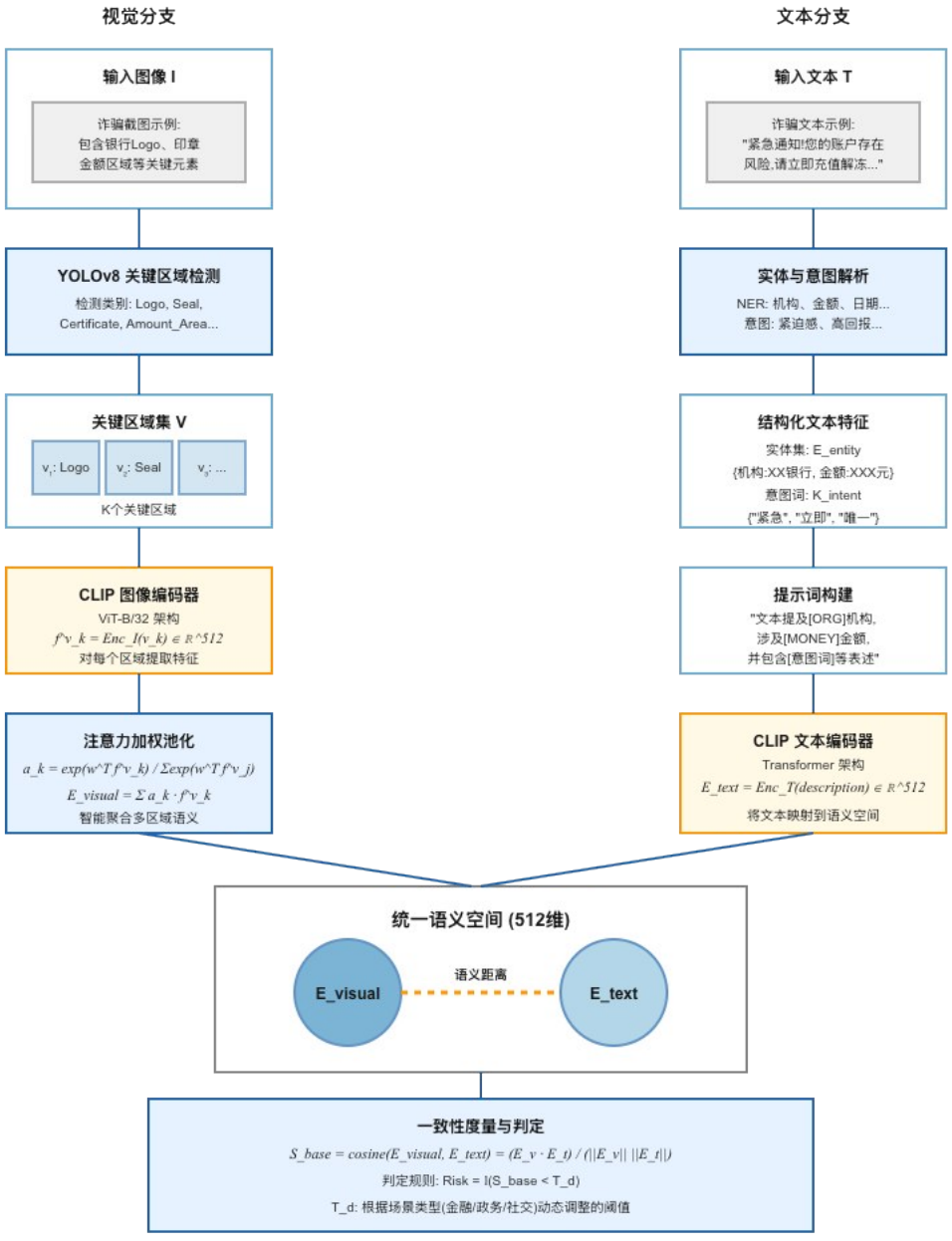

图3　基于CLIP的跨模态语义对齐与一致性度量流程

$$S_k = \frac{f_k^v \cdot E_{text}}{\| f_k^v \| \ \| E_{text} \|}, \quad Contradiction_k = \mathbb{I}(S_k < \eta) \tag{10}$$

其中 S_k 为第 k 个视觉区域与文本的余弦相似度，η 为矛盾阈值，$\mathbb{I}(\cdot)$为指示函数。

自适应阈值决策：固定的全局阈值 T 无法适应不同场景（如严肃的政务通知 vs. 营销海报）。为此，我们引入一个轻量级场景分类器，用于预判图文内容所属的领域 d ∈ {金融、政务、社交、电商}。该分类器采用多层感知机（MLP）架构，包含两个隐藏层（维度分别为256和128），以图文融合特征 $[E_{visual}, E_{text}]$ 作为输入，使用交叉熵损失函数进行训练。在验证集上，该场景分类器的平均分类准确率达到94.2%，证明了其有效性。

每个领域 d 关联一个动态阈值 T_d，该阈值在验证集上通过最大化F1分数确定。最终风险判定为：

$$Risk = \mathbb{I}(S_{base} < T_d) \quad (11)$$

4 实验与结果

4.1 数据集

目前，公开研究领域尚缺乏面向电信诈骗场景、标注图文语义一致性关系的大规模数据集。为此，本研究构建并公开了一个名为“Fraud-Bench”的电信诈骗图文检测数据集。该数据集的建设分为两个阶段：原始URL收集与筛选、图文对构建与语义标注。

（1）原始URL收集与筛选：我们相关单位提供的真实电信诈骗案例数据库中获取了约50万条涉嫌诈骗的网站链接（已上传至GitHub平台，详见数据可用性声明）。这些链接覆盖仿冒政务类、金融诈骗类、虚假营销类等多种诈骗类型，并保留了页面元数据（如标题、描述、域名等）。由于原始网页内容可能随时间变化或失效，我们采用分布式爬虫对仍可访问的链接进行了内容抓取，最终获得约18万组有效（图像，文本）对。

（2）图文对构建与语义标注：在上述有效图文对的基础上，我们进一步进行了质量控制与语义一致性标注。具体流程如下：

● 对每个网页，提取主体图像（如截图、伪造公文、Logo等）与关联文本（如标题、诱导性描述、金额等），形成图文对。

● 由三名具备信息安全背景的标注员独立判读，判断图文是否在核心事实、主张或意图上存在欺骗性冲突（“语义矛盾”）或逻辑自洽（“语义一致”）。

● 为量化标注员间一致性，引入标注一致性评估（使用Fleiss' Kappa）。经计算，三位标注员的Fleiss' Kappa系数为0.87，表明标注结果具有高度的一致性。

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \quad (13)$$

其中 $\bar{P}$ 为标注者之间实际一致的比例，$\bar{P}_e$ 为随机一致性期望。

● 最终构建的标注数据集包含12,500个高质

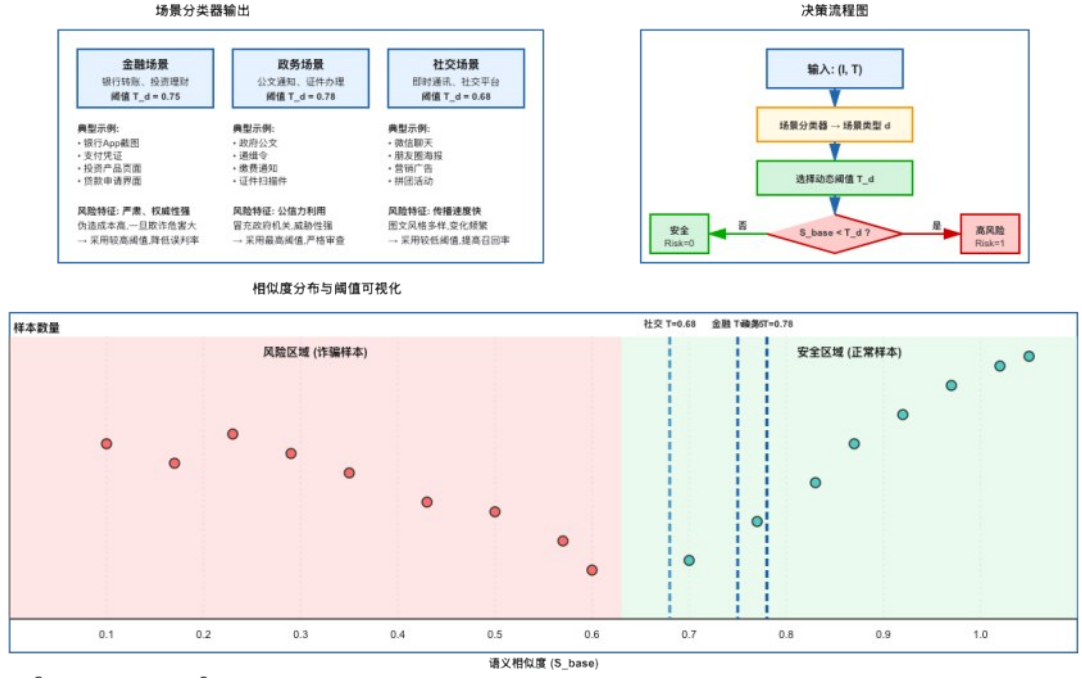

图4 场景自适应动态阈值机制与决策示意

量图文对，其中约40%来源于真实诈骗案例（经脱敏处理），30%来源于上述爬取内容中自然存在的矛盾样本，30%为从公开互联网收集的正常宣传物料（作为负样本）。数据按8:1:1划分为训练集、验证集和测试集。

需要说明的是，GitHub仓库中发布的是原始约50万条URL链接及其元数据，供研究者进行扩展采样与验证；而本论文实验所用的12,500个标注图文对，是基于该URL集合经过内容抓取、清洗与人工标注后构建的子集。研究者可通过仓库中的采样脚本和标注规范复现本实验。

4.2　评估指标与对比方法

为全面评估模型性能，我们采用以下四项常用分类指标：准确率（Accuracy）、精确率（Precision）、召回率（Recall）和F1分数（F1-Score）。

实验选取了七类具有代表性的基线方法进行对比：

1.Text-Only：仅基于文本模态。使用预训练的BERT模型对输入文本进行编码，并接分类层进行二分类（欺诈/正常）。

2.Image-Only：仅基于图像模态。使用在ImageNet上预训练的ResNet-50模型对输入图像进行编码并分类。

3.Early Fusion (ResNet+BERT)：早期特征融合方法。分别使用ResNet-50和BERT提取图像与文本的特征向量，将两者直接拼接后，输入到一个全连接层进行分类。

4.ViLBERT：经典的多模态预训练模型。在本数据集上对ViLBERT模型进行端到端的微调。

5.CLIP Zero-Shot：利用原始CLIP模型的零样本分类能力。通过手动设计提示词模板（如“A photo of [text description]”），计算图像与文本描述的匹配概率，并通过设定阈值将其转化为二分类结果。相关评估公式如下：

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{14}$$

$$Precision = \frac{TP}{TP+FP} \tag{15}$$

$$Recall = \frac{TP}{TP+FN} \tag{16}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{17}$$

其中TP, TN, FP, FN分别表示真正例、真负例、假正例、假负例的数量。

6.CLIP Fine-tuned：在本数据集上对CLIP模型进行端到端微调，作为衡量本文改进效果的强基线。

7.Qwen-VL (7B)：采用主流多模态大语言模型Qwen-VL，通过构建提示词（Prompt）“请判断给定的图文对在语义上是否一致，回答是或否”进行零样本推理，作为最新MLLM基线的代表。

4.3　实验结果与分析

各对比方法及本文方法在FraudBench测试集上的性能对比如表1所示。

表1　不同方法在FraudBench测试集上的性能对比（%）

方法	Accuracy	Precision	Recall	F1-Score
Text-Only	71.2	68.5	65.3	66.9
Image-Only	73.8	70.1	72.4	71.2
Early Fusion	78.5	75.6	76.8	76.2
ViLBERT	85.4	83.9	82.1	83
CLIP Zero-Shot	80.1	81.5	77	79.2
CLIP Fine-tuned	88.2	87.5	86.9	87.2
Qwen-VL (7B)	84.5	85.1	80.3	82.6
Ours	93.5	94.1	91.4	92.7

通过对表1数据与图5的分析可以看出，单模态方法（仅文本或仅图像）的性能相对最低，说明依赖单一信息源难以有效应对图文内容相互矛盾的复合欺诈场景。在此基础上，简单的特征拼接融合方法相比单模态性能有明显提升，但由于缺乏深度的跨模态特征对齐，其提升程度仍有限。而基于跨模态注意力机制的大规模预训练模

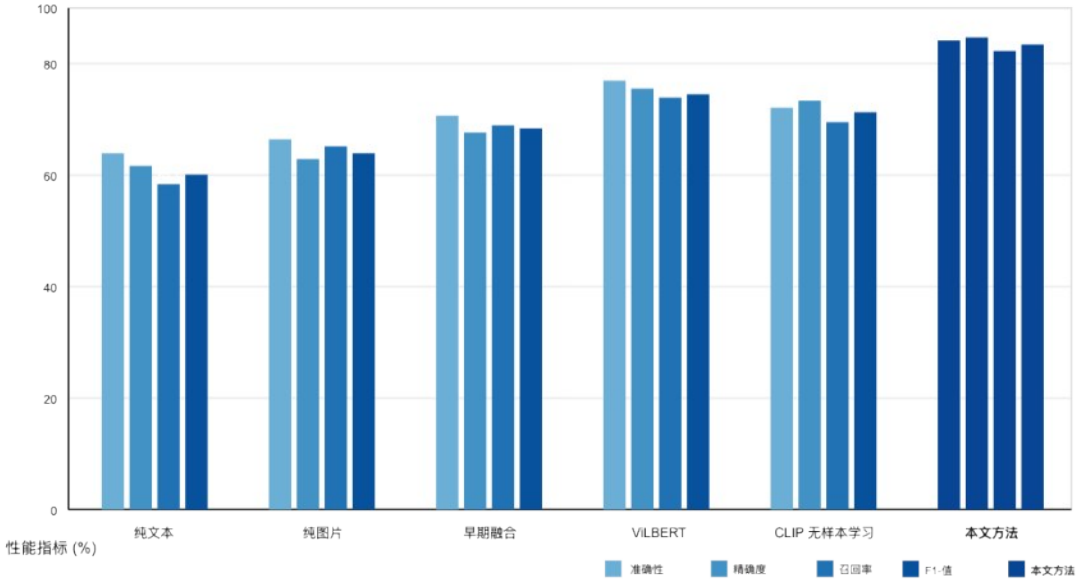

图5　不同方法在FraudBench测试集上的性能对比

型（如ViLBERT）虽然效果较好（F1值达83.0%），但模型结构复杂，计算开销较大。CLIP模型在零样本设定下表现出较强的跨模态匹配能力（F1值79.2%），但未针对欺诈场景进行优化。经过在本数据集上微调后，CLIP Fine-tuned基线性能提升至87.2%，展现了领域适配的重要性。值得注意的是，参数规模达70亿的Qwen-VL模型在零样本推理下取得了82.6%的F1值，展现了MLLM的强大潜力，但其性能仍低于经过专门微调的CLIP模型，且其推理延迟（约2.5秒/对）远高于本方法（420毫秒/对），难以满足实时检测需求。

本研究提出的方法在所有评价指标上均显著优于现有基线模型，最终取得了92.7%的F1分数。这一结果验证了所采用的“领域自适应关键区域检测”与“深度跨模态语义对齐”相结合策略的有效性。

4.3.1　分场景性能分析

为评估模型在不同诈骗场景下的表现，我们在测试集上对各子类分别进行了评估，结果如表2所示。

表2　本方法在不同场景下的性能对比（F1-Score, %）

场景	仿冒政务类	金融诈骗类	虚假营销类
F1-Score	94.1	91.8	92.3

结果表明，本方法在所有三个场景中均取得了均衡且优异的性能。其中，仿冒政务类场景的F1值最高（94.1%），分析认为，这是因为该类场景中的关键视觉元素（如国徽、公章）和文本实体（如“公安局”、“法院”）具有高度标准化的特征，模型更容易捕捉到图文间的语义矛盾。而金融诈骗类场景中，伪造的UI组件和金额数字形式多样，给检测带来一定挑战，导致性能略低。

4.3.2　误判案例分析

为进一步分析模型局限性，我们选取了测试集中部分误判样本进行人工审查。主要发现两类典型误判：

极度细粒度的语义矛盾：例如，一张伪造的银行转账截图，图像中的金额数字与文本中的金额完全一致，但转账流水号（极小的文本）与文本描述不符。由于当前YOLOv8模型未能有效检测到“流水号”这一微小区域，导致关键信息丢失，模型误判为一致。

需要外部知识推理的矛盾：例如，图像中是一个知名银行的Logo，文本承诺“年化收益50%”。虽然从常识判断这存在矛盾，但若模型未学习到“银行理财产品收益率有法定上限”这类外部知识，仅凭语义相似度计算可能无法判定其为矛盾。

这些案例为本文方法的进一步优化指明了方向，未来工作将探索引入OCR技术增强对图像内微细文本的提取，以及融合外部知识库来提升模型的推理能力。

4.3.3　关键参数敏感性分析

本方法中，关键区域筛选时的超参数 α 和 β 直接影响输入到CLIP模型的视觉信息质量。为分析其敏感性，我们固定β=0.5，在{0.3, 0.5, 0.7, 0.9} 范围内调整α，并在验证集上观察F1分数的变化。如图6所示，当α=0.7时模型性能最优。α过小（<0.5）会导致过度依赖先验知识，可能遗漏置信度低但实际重要的区域；α过大（>0.8）则使模型过分信任检测器的置信度，可能引入背景噪声。实验表明，模型性能在α∈[0.5, 0.8] 的范围内保持相对稳定，证明了本方法对超参数具有一定的鲁棒性。

4.4　消融实验

为验证本文方法中各个核心组件的贡献，我们设计了系统的消融实验，结果如表3所示。

表3　消融实验结果（F1-Score, %）

模型配置	F1-Score
完整模型（Ours）	92.7
w/o 关键区域检测（使用全图）	87.3
w/o CLIP（使用ResNet+BERT拼接）	76.2
w/o 动态阈值（使用固定阈值0.5）	90.1
w/o 微调YOLO（使用通用YOLO）	89.5

消融实验结果表明：

关键区域检测模块至关重要：移除该模块，直接编码整幅图像，导致F1值下降5.4%（从92.7%降至87.3%）。这证明了对关键视觉信息进行聚焦，避免背景噪声对语义稀释的必要性。

在分析各组件贡献时，可用公式量化组件移除带来的性能变化。如下：

$$\Delta_{comp} = F1_{full} - F1_{w/o\ comp} \tag{18}$$

$F1_{full}$ 为完整模型的F1值，$F1_{w/o\ comp}$ 为移除某组件后的F1值。

CLIP提供的深度语义对齐是核心：将CLIP编码替换为ResNet与BERT的简单特征拼接后，性能急剧下降16.5个百分点（F1 76.2%），凸显了在统一语义空间进行深度比对的优势，远胜于简单拼接的浅层融合。

动态阈值机制提升适应性：采用固定阈值（0.5）代替场景自适应的动态阈值，性能下降2.6%，说明动态机制能更好地应对不同诈骗子类型的差异性。

模型微调具有实际意义：使用通用目标检测权重而非在诈骗数据集上微调过的YOLOv8权重，F1值下降3.2%，表明面向特定场景的模型优化能有效提升对小目标和特定类别的检测精度。

4.5　鲁棒性测试

实际应用中，诈骗图像常会遭受质量退化或人为干扰以规避检测。为评估本方法的稳定性，我们在测试集图像上模拟了两种常见扰动：

高斯噪声：添加标准差σ=0.05的高斯噪声，模拟低质量传输或截图。

随机裁剪：随机裁剪图像面积的10%–20%，模拟对抗性遮挡。

我们以F1-Score为主要指标，并选取了依赖底层图像结构的感知哈希（pHash）方法作为对比基线。结果如表4所示。

表4　鲁棒性测试结果对比（F1-Score, %）

测试条件	本方法	pHash基线
干净图像	92.7	74.8
高斯噪声 (σ=0.05)	90.6	59.1
随机裁剪 (10–20%)	89.4	52.3

在描述扰动影响时，设计了性能衰减度量公式。如下：

$$\Delta_{robust} = F1_{clean} - F1_{perturbed} \quad (19)$$

$F1_{clean}$ 为在干净测试集上的 F1 值，$F1_{perturbed}$ 为在扰动图像上的 F1 值。

测试结果显示，在施加噪声和裁剪扰动后，本方法的 F1-Score 分别下降了 2.1% 和 3.3%，表现出较强的鲁棒性。相比之下，基于底层像素结构的 pHash 方法性能衰退超过 15%，对干扰极为敏感。这主要得益于本方法基于 CLIP 高层语义特征进行比对，其对局部的像素变化或缺失不敏感；同时，多区域检测与聚合机制也分散了单区域被破坏的风险。此外，在单张扰动图像上，本方法的平均处理时间仍保持在约 420 毫秒，满足实时性要求。

5 结束语

本文针对当前电信诈骗手段日趋视觉化、复合化，尤其是“图文并悖”新型欺诈模式带来的严峻挑战，提出了一种基于跨模态语义对齐的图文一致性检测方法。该方法创新性地构建了“关键区域检测—深度语义对齐”的技术框架。

首先，通过针对诈骗场景微调的 YOLOv8 模型，精准定位并提取图像中 Logo、印章等高欺诈相关性的视觉元素；同时，利用文本解析模块抽取出文本中的关键实体与欺诈意图。进而，借助在大规模图文对上预训练的 CLIP 模型，将上述视觉区域与文本信息编码至统一的语义空间，实现深度的跨模态语义对齐。最后，通过计算语义相似度并结合场景自适应的动态阈值机制，量化并判定图文之间的语义矛盾。在自建的“Fraud-Bench”电信诈骗 URL 数据集上的实验表明，本文方法取得了 92.7% 的 F1 值，显著优于多种单模态及多模态基线模型。系统的消融实验验证了关键区域检测、CLIP 深度语义对齐以及动态阈值等核心模块的有效性与必要性。此外，鲁棒性测试结果证明，该方法对图像噪声、随机裁剪等常见干扰具有良好的稳定性，且推理效率满足实际部署的实时性要求。

本研究主要聚焦于图文内容本身的语义一致性检测。为进一步提升系统的鲁棒性，未来工作将探索将本方法与图像篡改检测、域名信誉分析等其他维度的证据进行决策级融合（如公式 $R_{fusion} = w_1 S_{base} + w_2 S_{tamper} + w_3 S_{domain}$ 所示），以期在更复杂的对抗场景下实现更精准的综合风险判定。

参考文献：

[01] LI J, DANG J, WANG Y, et al. Image-Based Telecom Fraud Detection Method Using an Attention Convolutional Neural Network[J]. Entropy, 2025, 27(10): 1013.

[02] YANG J, LI S, HUANG Z, et al. An improve fraud detection framework via dynamic representations and adaptive frequency response filter[J]. Scientific Reports, 2025, 15(1): 19051.

[03] 梁飞, 张世星, 陈子睿. 基于威胁环境感知与大模型特征增强的区块链异常交易检测模型[J]. 数据与计算发展前沿(中英文), 2025, 7(6): 23-34.
LIANG F, ZHANG S X, CHEN Z R. Blockchain Anomaly Transaction Detection Model Based on Threat Environment Awareness and Large Model Feature Enhancement[J]. Frontiers of Data & Computing, 2025, 7(6): 23-34.

[04] 牟宇佯, 芦天亮, 陈亮. 电信网络诈骗犯罪中星链设备溯源方法[J]. 情报杂志, 2025, 44(4): 1-9.
MU Y P, LU T L, CHEN L. A Traceability Method for Starlink Devices in Telecom Network Fraud Crime[J]. Journal of Intelligence, 2025, 44(4): 1-9.

[05] 吕仁堃, 孙鹏, 郎宇博, 等. 面向深度伪造检测的高效自解释图神经网络[J]. 计算机应用研究, 2025, 42(6): 1832-1840.
LYU R K, SUN P, LANG Y B, et al. Efficient self-explaining graph neural network for deepfake detection[J]. Application Research of Computers, 2025, 42(6): 1832-1840.

[06] 刘晓龙, 刘欢, 赵耀, 等. AIGC 伪造内容被动检测与主动防御技术综述[J]. 中国科学: 信息科学, 2025, 55(9): 2250-2288.
LIU X L, LIU H, ZHAO Y, et al. Passive detection and active defense for AIGC-generated fake content: a survey[J]. Scientia Sinica (Informationis), 2025, 55(9): 2250-2288.

[07] 杨红梅, 赵勋. 人工智能赋能网络安全的挑战与应用[J]. 中兴通讯技术, 2025, 31(3): 39-43.

YANG H M, ZHAO X. Challenges and Applications of AI-Enabled Cybersecurity[J]. ZTE Technology Journal, 2025, 31 (3): 39-43.

[08] 刘治杰, 丁锰. 基于多模态特征融合的恶意程序分类研究[J]. 计算机应用与软件, 2025, 42(5): 311-319.
LIU Z J, DING M. Malware Classification Research Based on Multi-Modal Feature Fusion[J]. Computer Applications and Software, 2025, 42(5): 311-319.

[09] 高建新, 孙锦平, 蔡瑜坤, 等. 人工智能犯罪与我国对策研究[J]. 中国科学院院刊, 2025, 40(3): 408-418.
GAO J X, SUN J P, CAI Y K, et al. Artificial Intelligence Crime and China’s Countermeasures[J]. Bulletin of Chinese Academy of Sciences, 2025, 40(3): 408-418.

[10] 张玲玲, 黄务兰. 基于ChatGPT API和提示词工程的专利知识图谱构建[J]. 情报杂志, 2025, 44(3): 180-187.
ZHANG L L, HUANG W L. Patent Knowledge Graph Construction Based on ChatGPT API and Prompt Engineering[J]. Journal of Intelligence, 2025, 44(3): 180-187.

[11] 陈咏豪, 蔡满春, 张溢文, 等. 基于参数高效微调及双流网络的人脸伪造检测[J]. 计算机工程与应用, 2025, 61(10): 288-298.
CHEN Y H, CAI M C, ZHANG Y W, et al. Face Forgery Detection Based on Parameter-Efficient Fine-Tuning and Dual-Stream Network[J]. Computer Engineering and Applications, 2025, 61(10): 288-298.

[12] 游畅, 黄诚, 田璇, 等. 基于多维特征的涉诈网站检测与分类技术研究[J]. 四川大学学报(自然科学版), 2024, 61(4): 33-42.
YOU C, HUANG C, TIAN X, et al. Fraudulent Website Detection and Classification Technology Based on Multidimensional Features[J]. Journal of Sichuan University (Natural Science Edition), 2024, 61(4): 33-42.

[13] 周业勤, 邱莉榕, 张熙. 基于词典增强的电信诈骗文本线索词提取模型[J]. 东北师大学报(自然科学版), 2025, 57(3): 86-94.
ZHOU Y Q, QIU L R, ZHANG X. Telecom Fraud Text Clue Word Extraction Model Based on Dictionary Enhancement[J]. Journal of Northeast Normal University (Natural Science Edition), 2025, 57(3): 86-94.

[14] 陈傲, 白恩健, 吴赟, 等. 融合CNN与ViT的深度伪造人脸篡改视频检测方法[J]. 东华大学学报(自然科学版), 2025, 51 (6): 62-69.
CHEN A, BAI E J, WU Y, et al. Deepfake Face Tampering Video Detection Method Fusing CNN and ViT[J]. Journal of Donghua University (Natural Science Edition), 2025, 51(6): 62-69.

[15] 余聪, 李柏岩, 刘晓强. 基于深度学习的网页违规图片检测[J]. 现代计算机, 2022(13): 45-50.
YU C, LI B Y, LIU X Q. Webpage Non-Compliant Image Detection Based on Deep Learning[J]. Modern Computer, 2022 (13): 45-50.

[16] 梁枭杰. 基于免疫深度网络的网页信息违规检测研究[D]. 天津: 天津理工大学, 2025.
LIANG X J. Research on Web Information Violation Detection Based on Immune Deep Network[D]. Tianjin: Tianjin University of Technology, 2025.

[17] XIANGJUN K. Construction of Automatic Matching Recommendation System for Web Page Image Packaging Design Based on Constrained Clustering Algorithm[J]. Mobile Information Systems, 2022, 2022: 9706598.

[18] 司海平, 李阔, 李婷婷, 等. 基于提示对比学习的小样本电信诈骗文本分类方法研究[J]. 计算机应用与软件, 2025, 42(5): 77-84.
SI H P, LI K, LI T T, et al. Research on Few-Shot Telecom Fraud Text Classification Method Based on Prompt Contrastive Learning[J]. Computer Applications and Software, 2025, 42(5): 77-84.

[19] 斯彬洲, 孙海春, 吴越. 基于大语言模型和事件融合的电信诈骗事件风险分析[J]. 数据分析与知识发现, 2025, 9(7): 38-51.
SI B Z, SUN H C, WU Y. Telecom Fraud Event Risk Analysis Based on Large Language Model and Event Fusion[J]. Data Analysis and Knowledge Discovery, 2025, 9(7): 38-51.

[20] 庄华, 马忠红. 电信网络诈骗犯罪预警的失灵与优化[J]. 情报杂志, 2025, 44(2): 116-123.
ZHUANG H, MA Z H. Failure and Optimization of Early Warning for Telecom Network Fraud Crime[J]. Journal of Intelligence, 2025, 44(2): 116-123.

[21] 尹彦, 张红斌, 刘滨, 等. 网络安全态势感知中的威胁情报技术[J]. 河北科技大学学报, 2021, 42(2): 195-204.
YIN Y, ZHANG H B, LIU B, et al. Threat Intelligence Technology in Network Security Situation Awareness[J]. Journal of Hebei University of Science and Technology, 2021, 42(2): 195-204.

[22] 张溢文, 蔡满春, 陈咏豪, 等. 融合空间特征的多尺度深度伪造检测方法[J]. 计算机工程, 2024, 50(7): 240-250.
ZHANG Y W, CAI M C, CHEN Y H, et al. Multi-Scale Deepfake Detection Method Fusing Spatial Features[J]. Computer Engineering, 2024, 50(7): 240-250.

[23] 冯畅, 吴晓龙, 赵熠扬, 等. 生成式伪造语音安全问题与解决方案[J]. 信息安全研究, 2024, 10(2): 122-129.
FENG C, WU X L, ZHAO Y Y, et al. Security Issues and Solutions of Generative Fake Audio[J]. Journal of Information Secu-

rity Research, 2024, 10(2): 122-129.

[24] 李泽卿, 黄诚, 曾雨潼, 等. 基于知识图谱嵌入的涉诈网络链接补全和关键节点识别[J]. 四川大学学报（自然科学版）, 2024, 61(3): 44-54.

LI Z Q, HUANG C, ZENG Y T, et al. Fraudulent Network Link Completion and Key Node Identification Based on Knowledge Graph Embedding[J]. Journal of Sichuan University (Natural Science Edition), 2024, 61(3): 44-54.