E-Commerce Letters 电子商务评论, 2026, 15(5), 302-309
Published Online May 2026 in Hans. https://www.hanspub.org/journal/ecl
https://doi.org/10.12677/ecl.2026.155519

Hans汉斯

基于端－边－云协同的电商直播多模态实时内容审核研究

陈　茂

贵州大学大数据与信息工程学院，贵州　贵阳

收稿日期：2026年3月8日；录用日期：2026年3月20日；发布日期：2026年5月20日

摘　要

随着电商直播平台的快速发展，“内容生产－实时互动－即时交易”深度耦合的业务模式在显著提升用户参与度和交易转化效率的同时，也放大了内容安全与交易合规风险。针对传统中心化审核架构在高并发、弱网络环境以及隐私合规要求不断强化背景下面临的时延波动大、带宽与算力成本高、敏感数据暴露面广等问题，本文围绕电商直播多模态实时内容审核场景，提出了一种端－边－云协同的技术体系。论文从多模态内容审核、边缘计算协同机制以及轻量化模型部署等理论基础出发，构建了适用于直播连续流场景的“StreamID-Segment-Event”流式数据模型，设计了“L1快速过滤-L2边缘精检-L3云端复核”的分层检测框架，并结合视觉检测、音频识别、文本理解与跨模态一致性校验，实现对违规内容、虚假宣传和误导性营销等风险的实时识别。与此同时，本文提出基于知识蒸馏、量化、异构加速和弹性容器化部署的边缘推理优化方案，以提升系统在资源受限条件下的实时性、准确性与可扩展性。在工程治理层面，论文进一步设计了覆盖安全合规、审计留痕、人机协同复核与线上评估的闭环治理机制，并构建了涵盖实时性、准确性、成本与用户体验的KPI指标体系。研究表明，端－边－云协同架构能够有效降低端到端审核时延和全量回传压力，在兼顾识别精度、系统成本与治理合规性的基础上，为电商直播平台多模态实时内容审核提供了具有工程可行性和应用价值的解决方案。

关键词

电商直播，内容审核，端－边－云协同，边缘计算，多模态检测

Study on End-Edge-Cloud Collaborative Multimodal Real-Time Content Moderation for E-Commerce Live Streaming

Mao Chen

College of Big Data and Information Engineering, Guizhou University, Guiyang Guizhou

文章引用: 陈茂. 基于端-边-云协同的电商直播多模态实时内容审核研究[J]. 电子商务评论, 2026, 15(5): 302-309.
DOI: 10.12677/ecl.2026.155519

Received: March 8, 2026; accepted: March 20, 2026; published: May 20, 2026

Abstract

With the rapid development of e-commerce live-streaming platforms, the deeply coupled business paradigm of "content production, real-time interaction, and instant transaction" has significantly improved user engagement and transaction conversion efficiency, while simultaneously amplifying risks related to content safety and transaction compliance. Conventional centralized moderation architectures face substantial challenges in such scenarios, including large latency fluctuations, high bandwidth and computing costs, and broad exposure of sensitive data, especially under conditions of massive concurrency, weak network connectivity, and increasingly stringent privacy and compliance requirements. To address these issues, this paper proposes an end-edge-cloud collaborative technical framework for multimodal real-time content moderation in e-commerce live streaming. Building on the theoretical foundations of multimodal content moderation, edge computing collaboration, and lightweight model deployment, a streaming data model termed "StreamID-Segment-Event" is constructed for continuous live-streaming scenarios. In addition, a hierarchical detection framework consisting of "L1 fast filtering, L2 edge-side fine-grained detection, and L3 cloud-side review" is designed. By integrating visual detection, audio recognition, text understanding, and cross-modal consistency verification, the proposed framework enables real-time identification of risks such as non-compliant content, false advertising, and misleading marketing. Meanwhile, an edge inference optimization scheme based on knowledge distillation, quantization, heterogeneous acceleration, and elastic containerized deployment is developed to improve real-time performance, accuracy, and scalability under resource-constrained conditions. At the engineering governance level, a closed-loop governance mechanism is further established, covering security and compliance, audit logging, human-machine collaborative review, and online evaluation. Moreover, a KPI system is constructed to comprehensively assess timeliness, accuracy, cost, and user experience. The results indicate that the proposed end-edge-cloud collaborative architecture can effectively reduce end-to-end moderation latency and full-volume backhaul pressure. By balancing detection accuracy, system cost, and governance compliance, it provides a technically feasible and practically valuable solution for multimodal real-time content moderation on e-commerce live-streaming platforms.

Keywords

E-Commerce Live Streaming, Content Moderation, End-Edge-Cloud Collaboration, Edge Computing, Multimodal Detection

1. 引言

电商直播通过将“内容生产–实时互动–即时交易”耦合在同一业务链路中，实现了更强的沉浸式体验与更短的转化路径，因而在提升转化率与用户参与度方面具有显著优势。然而，这种强耦合机制也进一步放大了内容安全与交易合规风险：一方面，违规信息可借助弹幕互动与主播话术在短时间内快速扩散；另一方面，交易行为往往与内容呈现、价格表达和促销承诺紧密绑定，使得虚假宣传、夸大功效、侵权引流等风险更易在“边看边买”的即时决策中触发并产生更广泛的负面影响。与传统图文电商相比，直播场景的内容载体更为复杂，不仅包含主播视频画面与语音流，还叠加商品图文卡片、弹幕/评论文本、

标题与脚本、挂车链接及外部跳转等信息源，由此形成典型的多模态数据融合需求与强实时处理要求[1] [2]。在大促活动、热点事件或头部主播开播等高峰时段，平台通常需要在数百毫秒至数秒的“处置窗口”内完成风险识别、证据生成与干预决策，例如风险标签标注、流量降权、暂停/限流推流、拦截商品链接、触发人工复核与留痕审计等，以尽可能降低违规传播范围与潜在交易损失[3]。

当前业界较为常见的审核流程仍以“端侧采集–中心云推理–结果下发”的中心化架构为主，即将音视频与文本等数据集中回传至云端进行统一推理与策略决策，再将处置结果回传至直播间或业务系统执行干预[4]。该模式在跨地域网络抖动、海量并发推理负载以及隐私与数据合规要求持续强化的背景下暴露出明显瓶颈：第一，端到云的链路时延与云侧排队时延叠加，使端到端响应呈现强波动性，难以稳定保障秒级甚至亚秒级处置需求；第二，为提升识别精度而持续回传高清音视频片段会显著推高带宽占用与云侧存储/计算成本，进而影响系统可扩展性与单位治理成本；第三，数据集中化处理扩大了敏感数据暴露面与攻击面，在“最小必要、目的限定、可追溯审计”等治理原则下，将面临更高的合规与安全压力[5] [6]。

边缘计算通过将计算、存储与网络能力下沉至更靠近用户和内容生产的位置，如城市接入点(Point of Presence, POP)、运营商多接入边缘计算(Multi-access Edge Computing, MEC)节点、就近机房或区域边缘节点，能够在弱网、高并发与跨域链路复杂的条件下实现就近推理与本地闭环处置，从而降低端到端时延、减少高成本回传流量并提升系统韧性[7] [8]。与此同时，电子信息相关专业在视频编码与传输协议、音视频信号处理、嵌入式与边缘设备系统、网络通信与时延优化等方面具有方法与工程基础，可为直播审核系统的端侧采集与特征构建、链路调优与带宽控制以及边缘推理部署与异构加速提供直接支撑与实现路径。

基于上述背景，本文聚焦电商直播多模态实时内容审核问题，提出端–边–云协同的技术体系，并重点研究以下关键问题：1) 面向多模态与连续流特征，如何构建流式数据模型并设计分层检测机制以满足实时性与稳定性要求？2) 在边缘算力、能耗与资源受限条件下，如何实现高精度推理与弹性部署，同时兼顾高峰并发与服务可用性？3) 在隐私保护与合规治理框架下，如何实现可审计、可解释、可追溯的“检测–处置–复核–回溯”闭环，从而提升平台治理有效性与社会责任落实能力。相较于现有研究主要聚焦法律治理、中心云审核或单一场景工程实现，本文的改进与贡献主要体现在三个方面：其一，面向电商直播连续流场景，构建“StreamID-Segment-Event”流式数据模型，将多模态异步数据统一映射为可事件化处理的在线对象；其二，提出“L1 快速过滤-L2 边缘精检-L3 云端复核”的分层检测框架，并将视觉、语音、文本与跨模态一致性校验纳入统一协同链路；其三，在技术体系之外，将边缘推理优化、处置策略、人机协同复核与 KPI 评估结合起来，形成兼顾实时性、准确性、成本与合规性的工程治理闭环[9]-[11]。

2. 理论基础与相关研究

2.1. 多模态内容审核与实时推理

电商直播的内容风险呈现“多源输入、快速扩散、形态多变”的特点，因此审核往往需要同时处理视觉、音频与文本等多模态信息。具体而言，视觉侧检测主要面向低俗内容、违法标识、侵权商标及敏感物品等风险要素，开展目标级检测与场景级识别；音频识别侧重于口播关键词、辱骂与涉黄语音片段检测，并结合声学事件识别补充对非语义异常的感知；文本理解则聚焦弹幕、标题与商品描述中的违规表达、夸大宣传与诱导交易等语义风险；在此基础上，通过跨模态一致性校验可识别“口播承诺–画面展示–商品信息”不一致所引发的虚假宣传或误导性营销问题[1] [12]。由于直播内容连续产生且处置具有强时效性，离线批处理难以满足治理需求，工程上通常采用流式处理与片段化推理：将直播流按时间切分为短片段或滑动窗口，在窗口内进行特征聚合与风险判定，并结合复杂事件处理(Complex Event

Processing, CEP)实现多条件触发(如“关键词命中 + 画面敏感目标出现”)的实时联动；同时引入在线特征对齐机制以校正音视频与文本异步到达所引入的时序偏差，从而提升实时判定与处置的稳定性[13]。

2.2. 边缘计算与端–边–云协同

欧洲电信标准协会(European Telecommunications Standards Institute, ETSI)提出的多接入边缘计算(Multi-access Edge Computing, MEC)强调在接入网边缘提供计算、存储与网络能力，为低时延、高并发与高带宽敏感业务提供就近服务能力。在电商直播审核中，端–边–云协同的核心是形成“分层处理、策略可控、闭环可运维”的体系：端侧负责采集、编码与轻量预过滤，降低无效数据进入后链路；边缘侧承担高频推理、流式特征生成与本地处置，以保障毫秒到秒级响应；云侧负责全局训练、策略编排(规则/阈值/人审队列)、跨区域联动与审计留痕，实现治理的一致性与可追溯性[7] [8]。

从学术界看，相关研究主要沿三条路径展开：一是从平台治理、算法责任与用户治理角度讨论直播电商与社交平台的治理边界，强调平台在规则制定、风险预防与公私协同治理中的关键作用[14]；二是从广电与融媒体场景出发，研究人机协同智能审核系统、直播流智能审核控制系统等工程实践，验证了“规则引擎 + 智能识别 + 人工复核”的基本可行性[9] [10]；三是从多模态学习、流处理、边缘计算与轻量化部署等通用技术研究出发，关注审核准确率、在线处理能力与低时延部署问题。从工业界看，腾讯云已提供实时音视频内容安全审核能力，支持策略配置、自定义词库、回调与音视频多场景审查；阿里云则在直播全链路安全体系中集成智能审核与禁推流机制，强调对视频、音频、封面与标题等多类型内容的联动防护[15] [16]。

不过，现有研究与实践仍存在不足：其一，平台治理研究多聚焦法律责任、规则约束与用户治理，对适配电商直播高并发、低时延业务的系统架构讨论相对不足；其二，已有工程实践多面向融媒或中心云审核场景，通常集中于单平台、单链路或中心化部署，对跨节点协同、边缘侧弹性推理与分层处置机制覆盖不足；其三，现有通用技术研究虽讨论多模态检测、模型压缩与边缘计算，但较少将流式数据建模、跨模态一致性识别、工程治理闭环及 KPI 评估体系面向电商直播场景进行统一集成。基于此，本文面向电商直播业务，提出端–边–云协同架构，在流式数据模型、L1-L2-L3 分层检测、边缘推理优化和治理评估闭环等方面对现有工作进行了系统性整合与改进。

2.3. 轻量化模型与知识蒸馏

受限于边缘节点算力、能耗与资源配额，直接部署云端复杂模型往往难以满足实时推理的吞吐与时延要求，因此需要通过轻量化技术实现“可部署、可扩展、可维护”。常用方法主要包括模型压缩(如剪枝、量化)与知识蒸馏：模型压缩通过削减冗余参数并采用低比特计算以降低计算与存储开销；知识蒸馏则基于“教师–学生”框架将云端大模型的表征与判别能力迁移至轻量模型，从而在边缘侧实现更优的精度–时延权衡[17] [18]。针对视频流检测任务，为进一步降低计算量并提升稳定性，工程上常结合抽帧与关键帧选择减少冗余推理，并选用轻量级骨干网络完成高频检测；在风险升高或命中可疑线索时，再触发更高密度采样或上送云端复核，以形成“低成本筛查–高精度复核”的分层推理策略[19] [20]。这种“轻量模型 + 分层触发”的设计能够在满足实时治理要求的同时，控制边缘资源消耗并提升系统整体可扩展性。

3. 端–边–云协同架构与数据流设计

3.1. 分层架构

为兼顾电商直播审核的实时性、可扩展性与治理一致性，本文构建端–边–云三层协同架构。

1) 端侧(主播端/采集端)：承担音视频采样与编码、基础时序对齐及轻量级预过滤(如敏感词快速匹配、低成本视觉粗检等)，并上报片段元数据与低码率特征，以减少冗余数据进入后续处理链路。

2) 边缘侧(城市 POP、运营商 MEC、就近机房)：作为低时延处理核心，承担协议接入、流式特征生成、轻量多模态推理与复杂事件检测，并在命中风险时执行本地处置动作(如暂停/限流推流、降低码率、屏蔽弹幕、拦截商品链接等)，形成快速闭环。

3) 云侧：负责统一的模型训练与评测、策略编排(阈值/规则/人审队列)、跨边缘节点协同联动，以及审计留痕与治理分析，从而保障策略一致性与可追溯性[7] [8]。

该分层设计遵循“高频低时延下沉、低频高复杂上收”的原则，将时延敏感决策尽量前移至边缘，同时保留云端全局优化与治理能力。

3.2. 流式数据模型与处理管道

为适配直播连续流与多模态异步到达特性，本文将数据对象抽象为“StreamID-Segment-Event”三元组：其中，Segment 为固定时长片段(例如 1~2 秒)，包含抽样视频帧、音频片段与对应文本窗口；Event 表示对 Segment 或窗口聚合结果的判定输出，包括检测命中结果、处置动作以及人工复核结论等。在处理管道上，边缘侧采用“消息队列 + 流处理”的典型在线架构：端侧通过实时消息传输协议(Real-Time Messaging Protocol, RTMP)/HTTP 直播流协议(HTTP Live Streaming, HLS)推流，边缘节点执行片段化拉流与抽帧；多模态特征以消息形式进入流处理引擎，完成窗口聚合、跨模态对齐以及基于复杂事件处理(Complex Event Processing, CEP)的联动规则执行；对命中与可疑样本，仅上送事件摘要与必要证据片段至云端，用于样本回流、模型迭代与审计记录[13] [21]。该管道在保证实时响应的同时，能够通过窗口化与事件化机制降低全量回传与全量存储压力。

3.3. 网络与传输优化

在跨地域与弱网条件下，直播审核链路的时延与稳定性高度依赖传输与编码策略。传输层可引入快速 UDP 互联网连接协议(Quick UDP Internet Connections, QUIC)以增强丢包场景下的鲁棒性并提升多路复用效率；同时结合自适应码率(Adaptive Bitrate, ABR)与可伸缩视频编码思想，在不损害审核关键特征(如关键帧纹理、字幕区域清晰度、目标轮廓可辨性)的前提下，动态控制码率与分辨率，从而降低带宽占用与回传成本[22]。进一步地，在端–边–云协同架构下，可采用“特征上报优先、原始片段按需回传”策略：默认仅上传低成本特征与元数据；当边缘侧触发高风险判定或需要云端复核时，再选择性回传命中/可疑片段，以将高成本回传限制在必要范围内并提升整体链路效率。

4. 关键技术方法

4.1. 分层多模态检测框架

本文采用“L1 快速过滤-L2 边缘精检-L3 云端复核”的分层策略：L1 在端侧以规则与轻量模型快速剔除明显正常内容，降低边缘侧处理压力；L2 在边缘侧以轻量多模态模型进行高频检测，输出风险评分与关键证据；L3 在云端利用大模型/多任务模型对高风险片段复核并生成可解释依据，同时将复核结论回灌边缘以动态校准阈值与策略[17] [18]。

4.2. 视觉侧：轻量目标检测与违规场景识别

边缘视觉模型可采用轻量检测器识别敏感标识、违规动作及涉黄涉暴场景。为控制计算开销，可对直播帧进行自适应抽帧(画面变化或动作强烈时提高采样率)，并通过滑动窗口投票抑制瞬时抖动带来的误报[13]。针对侵权商标/商品图，可结合特征检索进行近似匹配，命中后触发更高等级复核[23]。

4.3. 音频侧：关键词与声学事件检测

音频审核主要包括自动语音识别(Automatic Speech Recognition, ASR)、关键词/敏感词识别与声学事件检测(如辱骂、尖叫、背景噪声突变等)。边缘侧可采用流式 ASR 或关键词唤醒模型实现低时延识别，并以声学事件检测补充“非语义”异常信号，从而提升对复杂场景的鲁棒性[24]。

4.4. 跨模态一致性与虚假宣传识别

针对“口播与画面/商品信息不一致”虚假宣传风险，可构建跨模态一致性判别：将 ASR 文本与商品标题/卖点进行语义匹配，同时对画面中的价格、字幕等光学字符识别(Optical Character Recognition, OCR)信息进行一致性校验；当一致性较低且伴随高转化诱导词时，提高风险等级并触发复核[25]。

4.5. 边缘推理优化：蒸馏、量化与异构加速

云端训练多模态大模型后，可蒸馏为边缘学生模型，并结合 8-bit 量化与算子融合提升吞吐；在图形处理器(Graphics Processing Unit, GPU)/神经网络处理单元(Neural Processing Unit, NPU)边缘节点上进行异构加速，并通过小批量微聚合降低推理开销，以支撑高并发实时审核[17] [18] [26]。

5. 工程实现与治理闭环

5.1. 弹性部署与运维

边缘侧审核服务采用容器化部署，并借助 Kubernetes 边缘扩展实现应用分发、弹性伸缩、灰度发布与快速回滚。全链路可观测性重点覆盖三类指标：推流质量(码率、丢包率、端到端时延)、推理性能(每秒查询数，Queries Per Second，QPS；第 95 百分位时延，P95；GPU/NPU 利用率)以及治理效果(命中率、误报率、处置时延)。命中结果进入工单与复核流程，形成“检测–处置–复核–回溯”的闭环管理机制。

5.2. 安全合规与可审计

内容审核涉及个人信息与敏感数据处理，应以零信任架构为基础开展设备身份认证与访问控制，关键数据采用端到端加密传输与存储；边缘节点遵循数据最小化原则，优先回传摘要信息与命中证据片段，降低集中化暴露面。云端保留审计日志，并对模型版本、阈值与策略变更进行版本化记录，以支撑合规审计、责任追溯与事后复盘[27]。

5.3. 人机协同与解释性

对高风险与争议样本应触发人工复核，以降低误杀并提升处置置信度。云端复核模块输出可解释证据(如触发片段、关键帧、关键词及一致性得分等)，增强治理透明度与可申诉性[10]。人工结论与申诉结果回流训练样本与规则库，用于持续迭代模型，并对边缘阈值与处置策略进行动态校准。

6. 评估指标体系与应用建议

6.1. KPI 定义

为全面衡量端–边–云协同审核体系的有效性，建议从实时性、准确性、成本与体验四个维度构建 KPI 集合：

1) 实时性：端到端审核时延(P95/P99)、处置闭环时延(从命中到执行干预/进入复核队列的时间)；
2) 准确性：精确率、召回率、F1，以及误报率/漏报率等风险识别质量指标；
3) 成本：回传带宽占用、边缘/云侧推理成本(算力资源消耗)、存储成本(证据留存与日志开销)；

4) 体验：误杀引发的投诉率与申诉率、直播中断率/限流率，以及对转化指标(点击、加购、成交等)的影响。

6.2. 评估方法

线上评估可采用灰度发布与 A/B 对照实验，对比端–边–云协同方案与云中心化方案在时延分位数(P95/P99)与回传带宽上的改善幅度，并同步观察命中率、误报率及业务转化等指标的变化。为校验识别质量，可对命中与未命中样本进行分层抽样，结合人工标注或复核结果估计误报与漏报水平。针对模型漂移与场景变化，可引入分布偏移指标(如特征分布变化)并配合在线监控与告警机制，及时触发阈值校准、回滚或再训练流程[3] [13]。

7. 结论与展望

本文面向电商直播场景下的多模态实时内容审核需求，构建了端–边–云协同的技术体系，并围绕流式数据模型、分层检测机制、边缘侧轻量化推理与工程治理闭环进行了系统化设计与论证。研究表明：通过将高频、时延敏感的检测与处置能力下沉至城市 POP/运营商 MEC 等边缘节点，并采用“StreamID-Segment-Event”的事件化建模与“消息队列 + 流处理”的在线管道，可在保证实时响应的同时降低全量回传压力；通过“L1 快速过滤-L2 边缘精检-L3 云端复核”的分层策略，以及蒸馏、量化与异构加速等边缘推理优化手段，可在资源受限条件下实现较优的精度–时延权衡，支撑高并发实时审核；在治理层面，零信任接入、端到端加密、数据最小化与审计留痕等机制能够提升系统的合规性与可追溯性，并与人工复核流程形成“检测–处置–复核–回溯”的闭环运行模式。此外，本文构建了覆盖实时性、准确性、成本与体验四维的 KPI 体系与线上评估方法，为端–边–云协同方案的工程落地与持续迭代提供了可操作的度量框架。

未来研究可在以下方向进一步拓展：

1) 面向生成式内容与深度伪造带来的新型风险，探索更鲁棒的检测方法与可信溯源机制；

2) 面向跨区域多边缘节点的协同治理需求，研究联邦学习与隐私计算在多方协作审核中的系统化融合与性能权衡；

3) 面向平台治理的长期目标，构建覆盖公平性、透明度与可解释性的评测体系，并推动内容治理指标与流程的标准化与可对标。

参考文献

[1] Kohavi, R., Tang, D. and Xu, Y. (2020) Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing. Cambridge University Press. https://doi.org/10.1017/9781108653985

[2] Cunningham, S. (2021) Causal Inference: The Mixtape. Yale University Press.

[3] Chandola, V., Banerjee, A. and Kumar, V. (2009) Anomaly Detection: A Survey. *ACM Computing Surveys*, **41**, 1-58. https://doi.org/10.1145/1541880.1541882

[4] OECD (2022) OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data. OECD Publishing. https://doi.org/10.1787/9789264196391-en

[5] NIST (2020) Zero Trust Architecture. NIST Special Publication 800-207.

[6] European Union (2016) General Data Protection Regulation (GDPR). European Union. https://eur-lex.europa.eu/eli/reg/2016/679/oj

[7] Shi, W., Cao, J., Zhang, Q., Li, Y. and Xu, L. (2016) Edge Computing: Vision and Challenges. *IEEE Internet of Things Journal*, **3**, 637-646. https://doi.org/10.1109/jiot.2016.2579198

[8] ETSI (2019) Multi-Access Edge Computing (MEC); Framework and Reference Architecture. ETSI GS MEC 003.

[9] 韩涛, 卜青原, 杨晓蕊. 云新闻发布平台直播流智能审核控制系统的设计与实现[J]. 广播与电视技术, 2023, 50(5): 55-58.

[10] 武开有. 基于人机协同的融媒内容智能审核系统探索与实践[J]. 广播与电视技术, 2025, 52(1): 41-45.

[11] 周辉, 魏日升. 直播电商治理的现实困境与优化路径[J]. 中国市场监管研究, 2025(10): 25-32.

[12] Tang, T., Wu, Y., Wu, Y., Yu, L. and Li, Y. (2022) Videomoderator: A Risk-Aware Framework for Multimodal Video Moderation in E-commerce. *IEEE Transactions on Visualization and Computer Graphics*, **28**, 846-856. https://doi.org/10.1109/tvcg.2021.3114781

[13] Carbone, P., Katsifodimos, A., Ewen, S., *et al*. (2015) Apache Flink™: Stream and Batch Processing in a Single Engine. *IEEE Data Engineering Bulletin*, **38**, 28-38.

[14] 晏青, 杜美玲. 驯顺与偏离: 社交媒体平台用户治理研究[J]. 新闻与传播研究, 2024, 31(1): 95-110, 128.

[15] 腾讯云. 实时音视频 内容安全审核[EB/OL]. https://cloud.tencent.com/document/product/647/77791, 2026-03-25.

[16] 阿里云. 直播全链路安全防护体系[EB/OL]. https://help.aliyun.com/zh/live/user-guide/security-overview/, 2026-03-25.

[17] Hinton, G., Vinyals, O. and Dean, J. (2015) Distilling the Knowledge in a Neural Network. arXiv: 1503.02531.

[18] Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., *et al*. (2018) Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 2704-2713. https://doi.org/10.1109/cvpr.2018.00286

[19] Redmon, J. and Farhadi, A. (2018) YOLOv3: An Incremental Improvement. arXiv: 1804.02767.

[20] 邵仁荣, 刘宇昂, 张伟, 王骏. 深度学习中知识蒸馏研究综述[J]. 计算机学报, 2022, 45(8): 1638-1673.

[21] Banks, J., *et al*. (2019) MQTT Version 5.0 Specification. OASIS Standard.

[22] Iyengar, J. and Thomson, M. (2021) QUIC: A UDP-Based Multiplexed and Secure Transport. IETF RFC 9000.

[23] Chen, W., Liu, Y., Wang, W., Bakker, E., Georgiou, T., Fieguth, P., Liu, L., and Lew, M. S. (2021) Deep Learning for Instance Retrieval: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 7270-7292.

[24] Graves, A., Mohamed, A. and Hinton, G. (2013) Speech Recognition with Deep Recurrent Neural Networks. 2013 *IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, 26-31 May 2013, 6645-6649. https://doi.org/10.1109/icassp.2013.6638947

[25] Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 1135-1144. https://doi.org/10.1145/2939672.2939778

[26] Howard, A.G., *et al*. (2017) MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv: 1704.04861.

[27] Dwork, C. and Roth, A. (2013) The Algorithmic Foundations of Differential Privacy. Now Publishers Inc. https://doi.org/10.1561/9781601988195