ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 41,No. 2,Mar. 2026,pp. 303－331
DOI:10. 16337/j. 1004-9037. 2026. 02. 003

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

以人为中心的可信视觉智能

高新波[1,2]，莫梦竟成[2]，张　灿[2]，袁　钰[2]，张明珠[2]，任路阳[2]，李　爽[2]，冷佳旭[2]

(1. 西安电子科技大学电子工程学院,西安 710071;2. 重庆邮电大学计算机科学与技术学院,重庆 400065)

摘　要: 本文围绕以人为中心的可信视觉智能,系统总结其应用现状、关键技术与发展趋势。随着计算机视觉从感知走向高自主决策与物理执行,视觉智能系统在隐私、公平、鲁棒、透明与安全等方面的风险日益突出,当系统输出可能影响人的安全与权益时,单纯追求性能已难以满足可信需求。为此,本文从计算机视觉视角梳理可信视觉智能的内涵与演进,强调人作为数据主体、认知参与者与最终控制者的多重角色,并提出以信息空间、认知空间与物理空间为主线的统一框架,构建“关注于人—服务于人—受控于人”的递进体系。围绕数据分析、模型设计与系统应用3个层面,本文总结公平与隐私约束下以人为对象的视觉数据分析方法,稳健且负责任的模型设计策略,以及以透明与安全为核心的人机协同控制机制,并结合图像增强、视频分析、机器人操作与三维视觉感知等场景进行分析。最后讨论了鲁棒评估、跨场景泛化、协同治理与可持续部署等挑战与研究方向,为真实世界可信视觉智能系统提供了路线图。

关键词: 可信视觉智能;以人为中心;计算机视觉

中图分类号: TP391.4　　**文献标志码:** A

引用格式: 高新波,莫梦竟成,张灿,等. 以人为中心的可信视觉智能[J]. 数据采集与处理,2026,41(2):303-331.
GAO Xinbo,MO Mengjingcheng,ZHANG Can, et al. Human-centered trustworthy visual intelligence[J].Journal of Data Acquisition and Processing,2026,41(2):303-331.

引　　言

人工智能作为一门模拟、延伸和扩展人类智能的新兴科学,核心目标在于让机器具备理解世界并服务人类的能力,而视觉感知不仅是人类获取信息的主要渠道,也是人工智能实现环境交互的基础。回溯生物进化史,自约5.4亿年前寒武纪的“光开关”被按下,动物开始具备以视觉区分“食物”与“敌人”的能力,视觉机制在生存竞争的驱动下不断重塑生命形态。受此启发,计算机视觉作为人工智能的核心分支,旨在赋予机器对视觉世界的感知与理解能力。其研究脉络从二维图像的特征表征与模式识别出发,逐步扩展到利用多视角信息重建三维场景,并在特征匹配与目标识别等关键任务上形成了一系列重要方法。随着算力、存储与数据规模的提升,视觉系统在复杂环境中的鲁棒性与泛化能力也随之显著增强。2010年以来,深度学习的快速普及为计算机视觉发展带来了革命性飞跃,卷积神经网络(Convolutional neural network, CNN)[1]、生成对抗网络(Generative adversarial network, GAN)[2]与Transformer[3]等架构推动了分类、检测与生成能力的质变。近年,随着视觉-语言模型(Visual language models, VLM)[4]、自监督学习[5]及扩散模型[6]的兴起,计算机视觉正加速突破“感知”边界,向着跨模态“理解与生成”的高阶智能演进。

基金项目: 国家科技重大专项(2025ZD0123601);国家自然科学基金(62472060);重庆市自然科学基金(CSTB2024NSCQ-QCX-MX0060, CSTB2023NSCQ-LZX0061,CSTB2023TIAD-STX0016)。
收稿日期: 2026-01-28;**修订日期:** 2026-02-28

随着计算机视觉技术从实验室走向社会深处，带来效率变革的同时，人工智能在伦理、道德与法律等方面引发的风险与挑战也日益突出。尤其当人工智能应用从内容推荐、语言识别等用户对错误容忍度较高的消费级场景[7]，走向自动驾驶、医疗诊断和金融决策等涉及人类切身利益的“深水区”[8]时，系统面临的要求已发生质变。在高影响场景中，关键不再仅仅是“给出结果”，而在于系统能否理解其所处场景并做出合规、安全的反应。若系统无法有效破解场景分析中的不确定性，其错误将不再局限于技术层面的性能失效，而是会在自动驾驶、医疗诊断等应用中被放大为对人的生命安全与财产利益的直接伤害，进而将会引发公众信任下降、技术质疑增加及治理成本上升等严重的社会损害。正因如此，“可信视觉智能”已成为未来视觉智能发展的必备要素。为避免上述链式后果，研究者与开发者必须将关注点从单一的精确性扩展至隐私性、公平性、稳健性、责任感、透明性与安全性等关键要素，在深入分析场景特定风险并遵循法律法规的前提下，通过配套的标准与伦理框架，确保人工智能以更安全、可控且可被信任的方式服务社会，以促成人类与智能系统的长期和谐共生。

实际上，人工智能的发展经历了多个阶段，已然蕴含了通向可信未来的演进脉络。如图1所示，以模型为中心的阶段，通过CNN[9]、GAN[10]和Transformer[11]等架构设计，人工智能在视觉识别与自然语言处理等领域取得了显著进展，标志着深度学习与生成式人工智能的崛起[12]。值得注意的是，因果人工智能与基于物理信息的人工智能在此阶段也逐渐受到重视，为提升模型的可解释性与可靠性提供了新的视角[13]。随着技术演进，人工智能逐渐进入以数据为中心的阶段，合成数据、数据治理与知识图谱等成为重点，这一阶段更强调数据质量与数据可信，通过减少歧视偏见、保护隐私来提升数据的可用性与合规性[14]。进一步，人工智能发展到以应用场景为中心的阶段，强调人工智能在自动驾驶、智能机器人等具体场景中的有效性与可控性，围绕降低决策不可预测性，将技术与现实世界复杂性深度结合，以更好地服务社会需求[15]。尽管模型、数据与场景导向的范式显著提升了性能与可用性，但目标函数往往仍以准确率、效率与可部署性为核心，难以充分回应公平、隐私、责任与可控性等与人直接相关的诉求。这促使人工智能从“技术最优”转向以人为中心的发展阶段，关注点从技术本身扩展到人类需求、数字道德与风险管理，并强调对信息滥用与数据违规的监管以确保技术产生积极影响[16-17]。最近，大模型的崛起为这一阶段注入了全新活力，通过学习并理解海量数据，显著提升了计算机视觉任务的效果，推动了图像分类、物体检测、语义分割以及视觉问答等任务的进步。在现代社会，以人为中心的设计理念强调在数据分析和模型训练过程中，如何更好地理解和服务于人。为了确保这种强大的技术应用能促进人类福祉，建立“人在回路中”(Human in the loop, HITL)与“人在回路上”(Human on the loop,

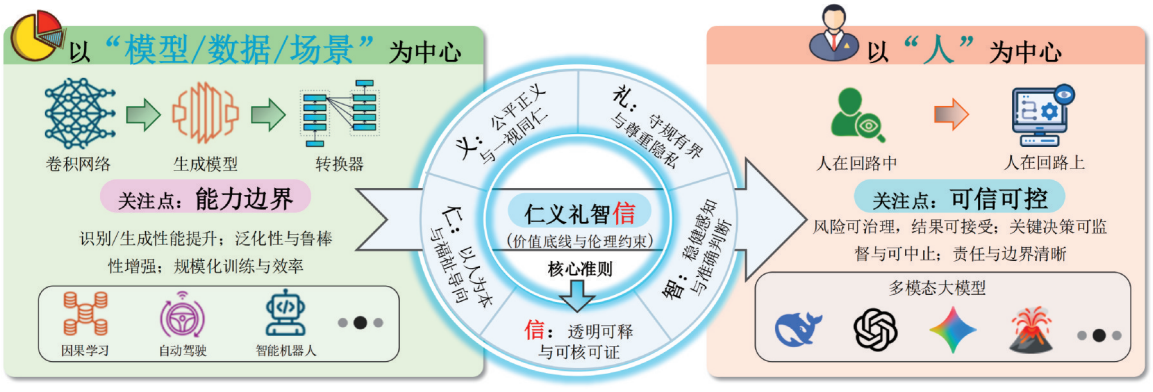


图1 从以“模型/数据/场景为中心”到“以人为中心”的可信视觉智能框架

Fig.1 Evolution from the “model/data/scenario-centered” to the “human-centered” framework for trustworthy visual intelligence

HOTL)的机制至关重要。人类必须始终保有对人工智能系统的控制权,包括在关键时刻中止其行为的能力,以最大限度降低潜在风险,从而在提升能力边界的同时提高信任与透明度,推动人工智能走向可持续且真正服务于全人类的发展范式。

随着人工智能向“以人为中心”的范式演进,技术与社会的深度交融使得“可信”超越了单纯的技术维度,成为行业发展的普遍共识与核心议题。这种共识的形成,源于人工智能在关键领域大规模应用时所暴露出的严峻挑战:从系统安全性面临的恶意攻击风险,到海量数据利用中的隐私泄露隐忧,再到算法偏见引发的社会不公以及“黑盒”决策带来的透明度缺失,这一系列风险叠加使得传统的性能导向难以为继[18-21]。因此,通过规范、标准与伦理框架对人工智能进行约束,已成为进入实际应用前必须满足的基本要求。在这一背景下,“可信”不再只是工程层面的可靠性指标,而是一套必须可执行的价值承诺:既要能力可靠、行为合规,也要回答技术将遵循何种伦理边界与福祉目标。为使这种价值承诺具备可落地的评估语言,国际学术界近年来逐步形成了较具共识的可信人工智能指标框架,代表性综述[22]将其概括为6项核心维度:人类监督与能动性、公平与非歧视、透明与可解释、鲁棒与准确、隐私与安全、以及问责机制。与此同时,中华传统文化“仁义礼智信”为上述维度提供了更具本土解释力的价值坐标:以“仁”统摄以人为本的社会福祉导向,以“义”衡量公平与非歧视,以“礼”界定隐私安全与数据使用边界,以“智”支撑鲁棒性与可靠决策能力,以“信”落实透明、可解释与问责机制。由此,传统价值不再停留于原则宣示,而能够以“可评估、可治理”的方式内嵌到技术设计与制度安排之中。早在2017年,何积丰院士便在香山科学会议上前瞻性地提出可信人工智能(Trustworthy AI)概念,强调可解释性、安全性与公平性等关键范畴,为后续治理体系奠定基础。近年来,伴随如《可信人工智能治理白皮书》[23]的发布及欧盟《人工智能法案》和中国《新一代人工智能发展规划》等全球政策的密集出台,可信人工智能已从理论讨论走向实践落地,成为学术攻关、产业探索与社会治理共同聚焦的核心议题。当前,无论是学术界对算法公平与透明性的基础研究,还是Microsoft、Google和Meta等产业界科技巨头对技术路径的探索,亦或是公众对隐私与责任边界的日益关切,都共同推动着可信人工智能成为全球范围内政策制定、技术攻关与社会治理的战略焦点,确立了其作为实现人机和谐共生必由之路的关键地位。

尽管可信已成为系统部署须满足的核心要求,且学术界已围绕其形成了涵盖鲁棒性、公平性、隐私保护、可解释性及问责审计等多维度的综述体系,并对评测指标与标准化难题进行了初步探讨[24-26],但现有的知识组织方式仍存在显著局限,难以满足“以人为中心”范式的深层需求。首先,既有综述多采用“以属性为中心”的并列式结构,倾向于对单一可信要素进行割裂讨论,缺乏对各维度在系统全生命周期中如何相互作用、相互制约的系统性分析。其次,在以人为中心的计算机视觉系统中,人类不仅是数据主体,更是认知参与者与最终控制者,深度嵌入系统运行的全过程;然而,现有工作鲜有系统刻画人在不同技术层级中的角色演化,也未能厘清其对可信机制设计的动态约束。此外,视觉系统的感知结果往往直接触发具体决策或物理行为,模型失效极易经由人类误判被放大为现实伤害,这一区别于语言或纯决策系统的独特风险传导机理,在现有综述中尚缺乏统一且清晰的结构化回应。鉴于此,本文基于计算机视觉视角,重构可信人工智能的分析框架,提出以“信息空间、认知空间和物理空间”为主线的以人为中心的可信视觉智能体系。如图2所示,该框架遵循“关注于人、服务于人、受控于人”的递进逻辑,建立起可信要素、系统层级与人类角色之间的精确映射关系,从而为不同视觉任务的模型设计与系统评估提供一套结构统一且可执行的研究路线图。

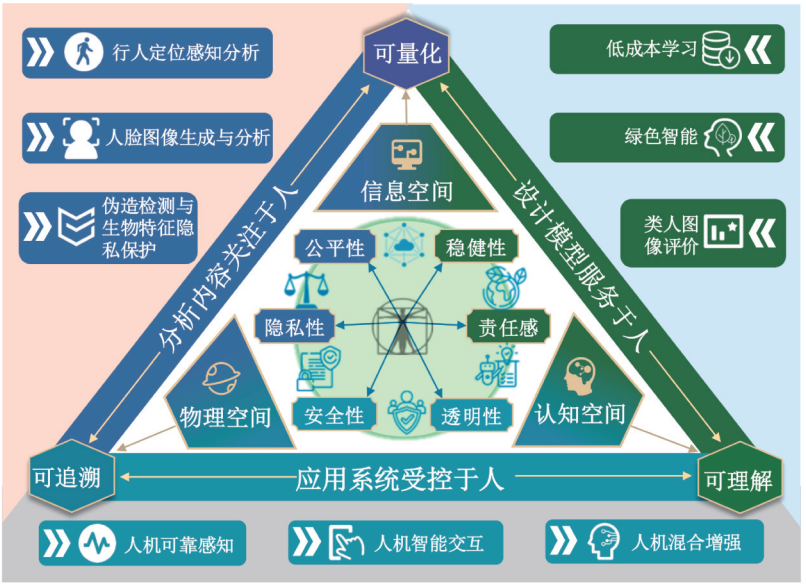


图2 以人为中心的可信视觉智能统一框架

Fig.2 Unified human-centered framework for trustworthy visual intelligence

1 以人为中心的可信视觉智能

随着计算机视觉技术由感知阶段向理解与生成阶段演进,其应用形态已从以实验室算法验证为主,逐步发展为嵌入现实系统的关键智能组件。在自动驾驶、公共安全、智能医疗与人机协作等高影响场景中,视觉系统的输出不再仅作为信息参考,而是直接参与现实决策与物理执行过程。在此背景下,视觉系统的潜在失效风险被显著放大,单纯依赖性能指标已难以保障系统在真实环境中的安全性与社会可接受性,亟需从系统整体机理层面对计算机视觉的可信性问题进行重新审视。

从系统运行机理出发,计算机视觉中的可信问题具有鲜明的人本特征。人既是视觉数据的来源与刻画对象,其隐私、差异性与潜在偏置贯穿数据获取与模型构建过程。同时,人也是视觉系统服务的对象与决策影响的承担者,模型推理结果将直接作用于人的行为、安全与权益。因此,可信视觉智能关注的不仅是算法性能的提升,还是如何刻画人,服务人,并规范视觉系统对人的影响方式。

基于上述认识,本文提出一种以人为中心的可信视觉智能框架,将视觉智能系统明确定位为分析内容关注于人、设计模型服务于人、应用系统受控于人的闭环系统,支撑视觉智能在复杂环境中的可靠运行,图2展示了框架结构。在此框架下,可信性被视为核心目标,由“公平性、隐私性、稳健性、责任感、安全性与透明性”6个基本内涵共同构成。为实现这一目标,本文进一步在信息空间、认知空间与物理空间这3个空间,对可信视觉智能进行系统建模,并通过“可量化、可追溯、可理解”的关系机制,将以人为中心的可信要求贯穿于模型、系统与应用的全流程。

在此基础上,本文对框架的3个关键维度分别展开分析。其中,分析内容关注于人位于信息空间,是整个体系的感知基石,其核心目标在于确立视觉智能在感知阶段应当算什么。为实现这一目标,以人为中心的通才模型需要在更小的视觉差异空间内进行稳定而细粒度的区分,以更充分地刻画人类外观结构、姿态形态与行为语义的多样性,并为后续认知与决策提供可靠表征。在这一层次中,研究对象以人像、行人、姿态与行为等细粒度人类视觉数据分析为切入,强调对外观结构线索、运动学规律与语义差异的联合建模。作为以人为中心的感知基石,该层次的可信性约束以公平性与隐私性为两条主线

并行嵌入数据建模过程。公平性要求模型在不同肤色、人种、性别与年龄等群体之间保持误差可比与稳定,避免在遮挡、低照度与拥挤等复杂条件下将漏检与误判系统性地集中到特定人群。隐私性则要求对生物特征与行为轨迹的处理遵循明确目的与最小必要原则,限制非必要的身份关联与画像化外溢,并使数据使用过程具备可审计、可追溯与可问责的治理属性。由此,分析内容关注于人能够在提升感知能力的同时守住公平与隐私边界,为可信视觉智能系统的构建奠定坚实基础。

设计模型服务于人位于认知空间,是整个体系的逻辑中枢。其核心目标是将人的需求与价值约束内化为模型的推理方式,从而回答视觉智能系统在认知与推理阶段应当怎么算的问题。为实现这一目标,模型不再仅被视为追求性能最优的映射函数,而是需要呈现可理解、可分析且可问责的认知行为,使其在复杂环境中以可预期的方式工作,让用户在关键场景下能够放心使用,并在风险与争议出现时具备可追溯、可治理与可问责的处置基础。为支撑这一服务于人的认知范式,模型可信性首先建立在稳健性这一能力前提之上。稳健性不仅意味着分布偏移、噪声监督与资源波动条件下的性能稳定,更要求模型对不确定性具备自我约束能力,避免在证据不足时输出过度自信的结论,并使失败模式可被及时识别与控制,保证结果的可靠性与一致性。在此基础上,模型进一步以责任感作为目标约束,使算法目标与评价机制能够反映真实世界的风险边界与社会规范,并在分布变化或部署受限时仍保持可执行与可持续,避免责任机制失效或流于形式。这样的可信演进路径以稳健性为支撑,确保系统可用性的基本底线,以责任感锚定价值与治理边界,最终形成模型增强人类能力而不替代主体的协同关系。由此,设计模型服务于人能够在认知层面确保视觉智能输出既可靠可用又可问责可治理,使技术发展始终服务于人类需求。

应用系统受控于人位于物理空间,是整个体系的控制终点与安全保障,其核心目标是确立人类在系统运行过程中的主导地位。为了实现这一目标,需要将人类持续纳入决策与执行的反馈闭环之中,即建立HITL与HOTL机制至关重要。为支撑这一受控范式,系统运行的可信性首先建立在透明性这一认知前提之上,通过将复杂的黑箱逻辑转化为认知对齐的决策证据链,使系统的行为逻辑对人类而言可解释,从而达成“人懂机器”的信任基础。在此基础上,系统进一步以安全性作为行为约束,使其在复杂环境与不确定条件下具备风险探测与状态感知能力,从而能够精准识别并响应人类的控制意图,实现“机器懂人”的深度协同。这种以透明性驱动认知对齐、以安全性保障行为约束的可信演进路径,最终汇变成“人机共生”的协同模式。通过赋予人类在关键时刻介入、修正或中止系统行为的决策权,视觉智能得以在发挥自主能力的同时有效规避失控风险。由此,应用系统受控于人能够保障技术在物理世界中的稳定运行,并使其在可信框架下最终回归于服务人类福祉的价值本位。

在上述框架中,“分析内容关注于人”“设计模型服务于人”与“应用系统受控于人”并非彼此独立的阶段,而是通过一组关键机制在信息空间、认知空间与物理空间中形成有机衔接的整体。三者共同构成以人为中心的可信视觉智能闭环,使系统能够从感知、认知到执行的全过程中持续对齐人类需求与价值。具体而言,从“关注于人”到“服务于人”,需要在信息空间中实现系统性能的“可量化”。模型的性能越好,越能够有效地服务于人类需求。性能的提升不仅体现在准确性和效率上,还应考虑其对人类行为和决策的积极影响。从“关注于人”到“受控于人”,需要在物理空间中实现推理过程的“可追溯”。权责的明晰促进了可信场景分析的健康可持续发展,使得用户能够在必要时追踪和理解系统的决策过程,确保对技术的掌控。无论是“服务于人”还是“受控于人”,都需要在认知空间中实现决策过程的“可理解”。只有当技术的工作原理和输出结果对用户来说是可理解的,才能建立起可信、可靠和可控的智能系统。这种理解不仅包括技术的运作方式,还涉及如何在复杂环境中与人类进行有效互动。

综上所述,本文构建了一个以人为中心的可信视觉智能统一框架,为视觉系统在复杂真实场景下

的设计、评估与协同控制提供全局视角。后续章节将以此框架为主线，分别从感知层面的分析内容关注于人、推理层面的设计模型服务于人和执行层面的应用系统受控于人这3个维度，系统梳理可信视觉智能的关键技术脉络与实践路径。

2 数据分析关注于人

在“以人为中心的可信视觉智能”体系中，围绕人类活动与个体生物特征的数据分析是系统理解现实世界的重要基础。相较于以性能指标为主导的传统视觉任务，“数据分析关注于人”将公平性与隐私性确立为与性能同等重要的核心约束，要求系统在刻画身份、行为与状态时避免对不同人群的错误与风险产生结构性偏向，并将个体信息的采集、关联与推断限定在必要范围内。

公平性与隐私性并非相互独立，而是共同构成“关注于人”的双重约束框架，其中公平性关心的是错误是否在群体之间分布得更均衡，例如不同肤色、人种、性别与年龄群体在误识率、漏识率等指标上不出现长期稳定的差距，并且在遮挡、低照度、拥挤与设备差异等更接近真实部署的条件下仍保持相对一致的表现。进一步地，公平性不仅体现在最终的对错上，也体现在系统输出的可信程度上，例如同样的置信度在不同群体上应具有相近含义，避免用统一阈值时让某些群体承担更高的误报或漏报风险。隐私性关心信息边界是否清楚并且能被执行。数据层面需要控制跨系统关联与长期积累带来的可追踪风险，避免把短期用途的数据演化为可识别、可定位和可画像的个人轨迹。模型层面需要注意表征被复用的风险，即使不保存原始图像，若身份特征可以在不同场景中被稳定匹配，也可能支持非授权的关联与追踪。生成式建模还引入了额外风险，如果模型过度记忆训练样本，可能在生成结果中泄露与原始个体高度相似的细节。因而，隐私性不仅要求少采集、少保存，也要求在表示、训练与部署阶段减少不必要的可链接信息，并为数据使用留下可追溯的记录与责任边界。

基于上述框架，本节以3项代表性任务勾勒“关注于人”的数据分析路径，具体框架如图3所示。行

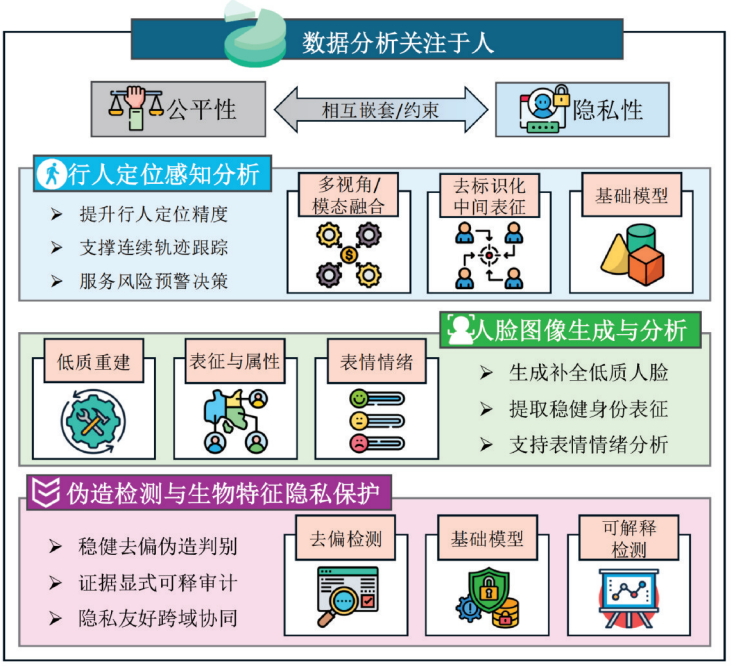


图3 数据分析关注于人

Fig.3 Human-focused data analysis

人定位与感知分析面向物理空间中的感知盲区,重点讨论拥挤与严重遮挡条件下的标签分配与特征学习,以提升个体感知的稳定性并减少对特定人群的系统性漏检,从而夯实公平性基础。人脸图像生成与分析聚焦身份信息的表征与关联,在分辨率受限、光照变化等条件下提升重建与识别鲁棒性,并推动群体间性能差异的收敛,避免身份识别在不同人群上的不均衡放大。伪造检测与生物特征隐私保护强调数据分析的安全与合规边界,针对生成式伪造带来的信任风险构建检测与溯源机制,并通过构建伪造检测与特征模糊机制,在不泄露敏感生物特征的前提下完成身份核验,实现数据利用与隐私保护的动态平衡。

通过以上3方面的系统阐述,本节旨在明确“关注于人”的数据分析目标,即以人的权利与处境为度量基准,在数据分析过程中实现公平性上的群体间误差均衡,并确立隐私性上的数据使用边界清晰且可治理,从而支撑可信视觉智能的可持续落地。

2.1 行人定位感知分析

行人定位与感知分析是“数据分析关注于人”的基础感知路径,贯穿存在性检测与定位、跨时空一致性关联以及行为与意图理解等环节。围绕这一链路,本节将公平与隐私作为贯穿性的任务取向,将相关研究归纳为3类相互衔接的技术抓手:(1)多视角与多模态感知增强在遮挡、低照度与拥挤条件下的稳定性,使漏检与误判不在特定人群上持续累积,更接近群体间误差均衡;(2)以姿态、骨架与几何等去标识化中间表征支撑行为理解,在满足功能需求的同时降低对可识别外观信息的依赖,使信息使用边界更清晰可控;(3)借助基础模型、泛化学习与弱监督机制降低对高敏感标注与身份关联数据的持续依赖,在跨场景可用性与数据使用可治理性之间形成更可持续的平衡。通过上述3条路径的递进组织,行人感知不仅关注能否稳定地看见与理解每一个体,也强调在公平与隐私约束下以更合理的方式实现这种能力。

面向复杂物理条件的定位与检测环节,多视角与多模态融合为公平性目标提供了关键支撑,其核心在于减弱环境差异对误差分布的结构性影响,使得各群体的感知质量不因条件差异而产生明显分化。多视角方法利用几何一致性缓解单视角遮挡与视角边缘带来的系统性漏检,文献[27-28]将空间互补引入检测与跟踪链路,通过全局融合获得更稳健的定位表征,从而降低拥挤与遮挡场景下漏检错误向特定人群集中累积的风险。多模态方法则通过跨模态不变性降低单一模态在特定环境下失效所引发的差异化误差,文献[29-30]通过因果去偏与动态模态组合促使模型学习可见光、红外与深度这3种模态间更加稳定的共享特征,使昼夜变化与光照不均条件下的性能波动不再由个别群体承担,从而更符合群体间误差均衡的公平诉求。

当行人感知从定位进一步走向行为与意图理解时,隐私性约束要求模型在表征层面降低对可识别外观信息的依赖,以更克制的信息完成必要推断,使信息边界能够内嵌于方法设计之中。基于姿态、骨架与几何的中间表征可以在不暴露人脸或细粒度纹理的情况下支持行为理解,文献[31-32]通过上下文对齐与异构骨架统一表明,关节点等运动学线索即可支撑高质量动作识别与迁移,从而减少对敏感生物特征的非必要调用。在人-物交互理解中,文献[33]以关键点引导区域建模,使判断依据更多落在动作结构与交互证据上,避免将身份外观与场景背景卷入推断,并为后续审计提供更清晰的归因线索。进一步地,文献[34]引入点云几何表征,在模态层面天然弱化外观细节,为隐私敏感场景提供可替代的分析通道。此类去标识化路线在强化信息边界的同时,也往往削弱了外观差异导致的表征偏置,与公平性目标形成一定协同。

在长尾场景与持续演化的应用需求下,统一泛化能力与弱监督学习成为兼顾公平与隐私的重要基础路径,其价值在于同时降低敏感数据依赖并缓解代表性不足带来的群体差异放大。更强的泛化底座

降低了围绕特定人群或特定场景反复采集高敏感数据与强标注数据的必要性,从而在数据源头缓解隐私合规压力,并降低跨系统关联驱动的画像外溢风险。文献[35-36]通过基础模型与多光谱通用表征提升跨姿态、跨尺度与跨任务迁移能力,使模型在新场景中更容易保持群体间误差的可比与稳定,避免因数据覆盖不足而对特定群体产生系统性失衡。面向无监督与跨模态设定,文献[37-38]在减少人工配对标注与身份关联数据需求的同时提升身份表征稳定性,使系统在较少依赖敏感关联数据的条件下仍具备可用性,并为公平性评估提供更一致的性能基础。

综上,行人定位与感知分析作为“关注于人”的基础链路,其目标不再仅仅是提升总体性能,而是在公平与隐私两条尺度下获得可检验的可信性。公平性方面,相关研究通过增强遮挡、低照度与拥挤条件下的稳定感知,使关键误差在不同人群间更可比、更均衡,从而抑制风险向特定群体集中。隐私性方面,去标识化表征与泛化学习共同降低了对可识别外观与高敏感关联数据的依赖,使数据使用边界更清晰且可治理。

2.2 人脸图像生成与分析

人脸图像生成与分析位于信息空间中对个体身份与心理状态进行表征与推断的关键链路,覆盖低质观测下的重建补全、身份一致性建模以及表情与情感理解等环节。围绕这一路径,相关研究逐步形成了以公平与隐私贯穿的方法取向,既要求不同肤色、性别与年龄等群体在重建与识别误差上趋于均衡、不因条件差异而显著分化,也强调在人脸这一高敏生物特征上减少不必要的信息调用与长期积累,使数据使用边界更清晰可治理。现有工作大体可以归纳为3类相互衔接的技术抓手,一类依托生成式先验与不确定性表达提升退化条件下的稳定性,使错误不易在特定群体上集中,同时降低对高敏真值监督的依赖,一类围绕身份保持与鲁棒表征提升跨条件一致性,使决策依据更多落在任务所需线索上并抑制可关联信息的过度外溢,另一类通过去噪净化、弱监督与语境建模治理主观偏差与标注噪声,使心理信号分析在保持可用性的同时更节制地使用数据。

在重建与补全环节,生成式先验被用于缓解遮挡、低分辨率与复杂光照带来的证据缺失,其关键价值并不止于提升视觉逼真度,更在于减弱退化条件对误差分布的结构性扰动,使不同人群在不利条件下的重建质量趋于均衡,不因退化程度差异而产生明显分化。遮挡与低分辨率往往诱发模型对少量可见纹理形成的捷径依赖,从而使重建误差更容易在成像更困难或数据覆盖不足的人群上累积。文献[39]将遮挡表情重新建模为多模态分布输出,通过显式表达不确定性抑制信息不足时的过度自信判断,为跨人群一致性评估提供更稳健的比较基础。与此同时,文献[40-41]通过弱真值或合成驱动降低对高精度纹理与扫描真值的依赖,使模型更容易覆盖长尾样本,从源头缓解代表性不足引发的群体差异放大,并相应减少高敏采集与密集标注所带来的隐私暴露面。

沿着身份一致性这一主线,研究关注点从补全像素转向维持可迁移的身份表征稳定性,进而把公平要求具体化为跨条件、跨人群的身份保持能力是否同等可靠。文献[42]通过身份条件化的潜变量扩散建模提升纹理生成的一致性,使身份线索在姿态、光照与分辨率变化下更稳定可控,从而降低差异化失败在特定群体上的集中风险。不过,身份保持能力越强,表征的可关联潜能也越强,隐私风险往往从原始数据是否被直接暴露转向表征是否被跨场景复用与长期积累。因此,这一方向除了追求一致性本身,还需要强调用途限定与访问边界,使身份一致性服务于授权场景下的必要生成与比对,避免演化为跨系统追踪与画像推断的通用接口。例如,文献[43]通过身份一致性约束的扩散生成框架在保持可控一致性的同时提升合成样本的可用性,使“以合成数据支撑授权场景下的必要比对/增广”成为可行路径,从而在一定程度上缓解真实人脸数据长期积累带来的隐私暴露风险。由此,身份一致性不仅是性能维度的改进点,也成为在群体误差可比与信息边界可控之间必须被审计与治理的关键能力。

分析转向表情与情感等心理信号时,公平问题更常表现为主观标注与文化语境差异导致的结构性

偏移,隐私问题则集中于高敏心理线索的采集、标注与推断是否过度。表情标签并非纯客观量,表达习惯与文化背景差异会使标注噪声在群体间呈现不对称分布,进而导致模型误判更易向某些群体集中。文献[44]将噪声分解为样本不可用与标注偏差并进行双阶段净化,使训练过程更不易固化主观误差,稳健一致性学习[45]则进一步提升噪声标注下的稳定学习能力,使输出在不同群体间更可比、更一致。面向微表情等更高敏感度场景,文献[46]以自监督机制减少对顶点帧等强监督标注的依赖,在降低采集与标注强度的同时减少心理信号被过度暴露的风险。文献[47]将情绪理解置于语境中建模,避免脱离上下文的冒进式解读,使系统在完成必要推断时更克制地使用信息,从而使数据使用边界更清晰、风险更可控。

综上,人脸图像生成与分析作为信息空间中“关注于人”的关键链路,其可信目标不再止于平均重建质量或识别精度的提升,而是要在公平与隐私两条尺度下获得可检验的稳定性。生成式先验、不确定性表达与稳健学习共同提升退化条件下的跨群体可比性,抑制错误向特定人群集中;弱真值依赖、弱监督与语境化建模降低对高敏采集与密集标注的持续依赖,并推动用途边界与访问边界内嵌进身份与心理信号分析过程,使数据使用更清晰且可治理。

2.3 伪造检测与生物特征隐私保护

伪造检测与生物特征隐私保护关注的是信息空间中的“信任”问题:在深度合成和物理呈现攻击不断演化的情况下,系统需要稳定地区分真实与伪造,并给出可追溯、可解释的安全决策。它不同于一般的分类任务,一旦判断失误,往往会直接带来安全风险;而且在设备质量、环境变化和人群差异的影响下,误报与漏报可能在不同群体之间不均衡,导致防护能力并非对所有人同等可靠。另一方面,为了提升鲁棒性,反伪造系统常常依赖大量真实人脸和攻击样本进行训练与更新,这又会带来隐私暴露和合规压力,限制系统的部署范围与长期迭代能力。围绕这些矛盾,现有研究大致形成3条逐步深入的路线:(1)强调更稳健、去偏的判别学习;(2)强调把模型依据的证据显式化,提高可解释性与可审计性;(3)在数据最小化的约束下,探索跨域部署与协同更新机制。总体而言,它们共同指向一种更以人为中心的防御形态:既让安全能力在不同人群间更一致、更可获得,也让数据使用的边界更清晰、更可控。

伪造检测面临的主要风险在于捷径依赖,检测器可能过度依赖某类频谱伪影、压缩纹理或局部光照模式,当这些线索与设备条件和人群外观分布发生纠缠时,就会出现对某些群体更高的误报或对某些场景更高的漏报。Liu等[48]首次面向人脸伪造检测的种族公平性构建公平伪造监测(Fair forgery detection,FairFD)评测基准并系统验证主流检测器的种族偏差,同时提出改进的公平性度量与无需重训练的偏差缓解方法以支持更可信的跨人群部署。文献[49]通过一致性驱动的频率去偏削弱对特定频域伪影的依赖,提升跨数据集与跨域泛化的稳定性,使防御失败不再由少数场景或少数群体承担。文献[50]进一步将人口统计学特征与伪造痕迹解耦,促使检测器仅基于伪造线索做出决策,从机制上降低跨种族与跨性别的性能差异。面向物理活体检测,多模态条件下的模态不可靠同样会带来差异化误差,文献[51-52]通过不确定性引导的抑制与对齐策略提升复杂光照与异构设备下的稳健性,使系统在多样环境中更接近一致的安全服务。此类去偏与稳健路线在改善群体可比性的同时,也减少将人群属性卷入决策的空间,使判别过程更少依赖与身份外观高度耦合的敏感线索。

为了让安全结论可核查、可问责,研究开始强调证据的显式化与可解释性,使性能差异能够被追溯到具体证据类型与决策依据,从而支持对误报与漏报来源的审计,也降低对可识别细节的无谓调用。文献[53]通过局部与全局伪造线索联合建模强化证据表达,使判断依据更集中于伪造痕迹本身;文献[54]引入视觉语言对齐增强语义解释能力,使检测过程更透明、证据更可追溯。证据链更清晰时,偏置分析不必停留在结果层面,而可以进一步检视模型究竟依赖了哪些线索,从而更有效地减少由外观相关项引发的群体差异。与此同时,证据导向的设计也促使系统把信息使用收敛到任务必要的伪造证

据上,减少对可识别纹理与身份细节的过度提取,使隐私边界更易被明确和治理。

在跨域部署与协同更新方面,隐私友好范式的意义不仅在于减少数据流转风险,也在于避免因合规与共享成本差异造成防护覆盖的不均衡,使安全能力更普惠地触达不同机构与不同用户。传统集中式训练依赖汇聚海量真实人脸与攻击样本,既触及合规风险也限制跨机构覆盖。文献[55]通过源自由域适配在不传输源域敏感数据的前提下完成部署,使防御能力能够以更低的数据共享成本覆盖更多场景。文献[56]以扩散生成方式补足跨域攻击样本分布,在减少真实攻击数据采集需求的同时提升对长尾攻击的覆盖,更贴近数据最小化原则。文献[57]通过联邦学习实现多方协同防御,在不出域传输原始数据的情况下共享模型能力,使隐私边界内的协作成为可能,并为持续演化的攻击形态提供更加可持续的更新机制。由此,反伪造能力不再依赖敏感数据的集中化积累,而是在隐私可控前提下实现更广覆盖与可持续迭代。

综上,伪造检测与生物特征隐私保护作为信息空间中“关注于人”的安全链路,关键不在于单点指标的提升,而在于形成对不同人群更一致可得的安全边界,并在数据最小化与可追溯治理下维持可部署、可更新的防御体系。稳健去偏检测提升跨设备、跨环境与跨人群的稳定防护,使误报与漏报更可比、更均衡;证据显式化与隐私友好部署减少了对可识别细节与集中式敏感数据汇聚的依赖,使数据使用边界更清晰且可治理。

3 模型设计服务于人

在“以人为中心的可信视觉智能”体系中,模型设计并不是对数据与算法的简单拼接,而是决定系统能否在真实世界中稳定可用,并以符合人类价值的方式输出结果的关键环节。与“数据分析关注于人”更侧重公平与隐私不同,“模型设计服务于人”强调把人的需求作为设计的出发点与落脚点:让系统在复杂环境中可预期地工作,让用户在关键场景下能放心使用、用得起、用得上,并在出现风险与争议时可追溯、可治理、可问责。本节不以输入数据类型限定讨论范围而以人的安全权益与使用体验作为模型设计的共同目标。

据此,本节将“稳健性”与“责任感”作为模型设计的两条相互嵌套的主线来展开,但不再将二者在模型设计中割裂为彼此独立的属性。这里的稳健性不仅指分布偏移、噪声监督与资源波动条件下的性能稳定,还包括工程与治理层面的稳健要求,例如在不确定性升高时不过度自信、在失败发生时能够被及时发现并受到控制。责任感同样不应仅被理解为评价维度更贴近人类偏好,而应强调评价与约束机制在真实条件下的可执行性与可持续性,避免在分布变化或算力紧张时出现责任机制“失效”或“形同虚设”。

基于这一理解,本节用3项关键策略共同刻画“服务于人”的模型设计路径,具体框架如图4所示。低成本学习通过降低对高强度标注与频繁数据更新的依赖,减少人力与资源投入,同时引入置信度校准与风险约束,使训练过程在降本前提下保持稳健,并抑制冒进式伪监督带来的误判风险。绿色智能将算力与能耗视为模型能力边界的硬约束,确保端侧与在线部署中维持稳定吞吐,并在资源受限时实现可预期的性能退化,从而以更低环境与经济成本支撑可持续落地。类人图像评价将人类感知、偏好与情绪体验转化为可计算的筛查信号与审计指标,为生成、增强与编辑等高风险任务提供了可解释的把关依据,并要求该把关机制在分布变化与生成失真条件下仍保持稳定有效。

通过以上3条路径,本节将“服务于人”落实为4类可操作目标:减轻负担(减少标注工作量与能耗开销)、降低风险(低把握结果提示风险并交由复核,必要时直接拦截或降级输出)、提升体验(依据感知质量与偏好评价进行优化与筛选)和支持追责(评测可复现、过程可记录、结果可审计)。

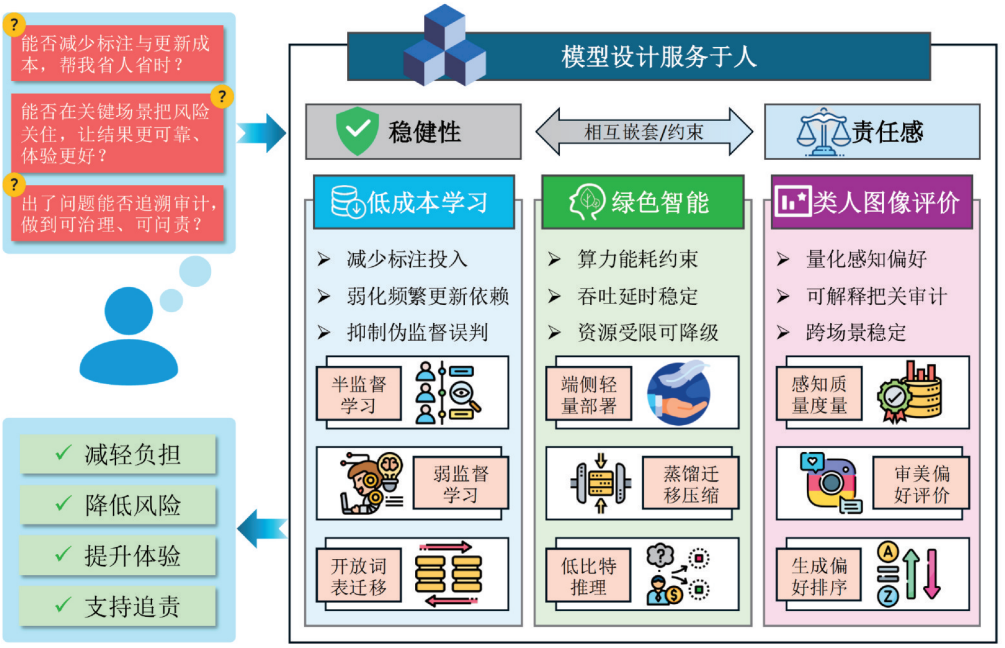


图4 模型设计服务于人

Fig.4 Human-serving model design

3.1 低成本学习

在可信视觉智能的学习过程中,标注既是成本瓶颈,也可能成为可靠性风险源。一方面,高代价标注限制数据覆盖面与场景多样性,导致模型更容易在真实分布中失灵;另一方面,稀缺标注与噪声标注会诱发过拟合、偏置放大与训练不稳定,进而削弱关键场景中的可控性与可解释性。因此,低成本学习不宜被狭义等同于“省标注”,更应被理解为一种面向可信落地的训练范式:在减少人力投入的同时,使模型在信息不充分时保持谨慎,并将不确定性与风险显式纳入优化与决策流程,从而实现“低成本与高可信”的协同。

从稳健性角度看,低成本学习强调在不完美监督下保持稳定学习。半监督、弱监督与自监督可利用一致性约束与结构先验缓解噪声标签与长尾样本带来的不稳定;跨域、跨视角对齐与开放集设定有助于提升分布偏移条件下的泛化能力;结合不确定性估计、置信度校准与难例挖掘,可在“数据有限、监督偏弱、场景持续变化”的现实条件下维持更清晰的性能边界。更关键的是,低成本学习在责任感层面必须避免“伪标签冒进”与“弱监督自我强化偏差”等风险。当模型无法识别自身不确定性时,节省的标注成本可能以更高的误判代价在部署阶段集中回流。因此,低成本学习需要将“何时采信模型输出、何时追加监督、何时触发人工复核”纳入训练与部署的一体化目标,使稳健性要求能够直接约束风险外溢。

半监督学习以“少量标注数据+大量未标注数据”为基本范式。文献[58]通过“弱增强生成高置信伪标签、强增强施加一致性约束”的组合取得显著效果。但伪标签机制天然面临确认偏差与噪声自强化风险,当阈值过严会造成未标注数据利用不足,且阈值过松会引入系统性噪声并破坏鲁棒性。围绕伪标签“数量-质量”权衡以及类别学习进度差异,文献[59-61]分别从课程式伪标签、自适应阈值与连续型软权重等角度调控伪监督强度,在提升未标注利用率的同时抑制错误累积,使训练增益更稳定、失败

模式更可控,从而更契合可信训练对稳健性与责任约束的双重要求。在半监督目标检测中,定位与分类耦合、类不均衡与伪框偏置更易导致偏差传播。文献[62]强调对伪标签偏置的抑制与类均衡建模,文献[63]通过“软教师”权重化未标注损失与框扰动筛选提升伪标签可靠性,使框架在更强检测器设置下仍能保持稳定收益。总体而言,这类方法将“节省标注”具体化为“在不确定条件下不过度自信”的训练机制,并通过自适应约束、校准与偏置抑制,使学习过程更稳健且更具风险可控性。

在弱监督学习中,部分标签学习为解决“标签不确定/不可靠”问题提供了一个关键的理论框架。面向可信性,关键不在于堆叠复杂结构,而在于明确在何种损失形式与噪声条件下仍能保持鲁棒性,并形成可解释的行为边界。文献[64]对平均型部分标签损失的鲁棒性进行系统刻画,指出有界损失与平均策略在特定设定下具有更稳定的优化性质,从理论层面为“稳健训练如何对应风险边界”提供了可解释支撑。与此同时,文献[65]基准从更现实的数据收集与统一评测协议出发,强调模型选择、协议一致与可复现对可落地可信性的基础作用。若评测协议不统一或复现不可得,即使指标提升,也难以为真实部署提供可靠承诺,更难支撑审计与问责,进而体现出责任感对稳健性的反向约束。

在更开放的泛化需求下,零样本与开放词表能力可在新概念出现时减少“逐类重标注”的依赖,使低成本学习扩展为面向持续演化的可扩展能力。例如,文献[66]通过视觉-语言知识蒸馏将开放词表识别能力迁移到检测框架,为“以文本扩展类别”提供低成本路径;文献[67]利用图像级监督扩展检测词表规模并增强长尾覆盖,体现弱监督数据更易获得的现实优势。在此类范式中,“服务于人”的价值不仅体现在节省标注劳力,还体现在对新需求的快速响应与对长尾场景的覆盖扩展。同时,开放词表带来的语义歧义与开放集误检同样需要被纳入风险治理框架,通过不确定性提示与复核触发机制降低误用风险。

综上,低成本学习通过降低重复标注与频繁数据更新的负担,并借助阈值自适应、软权重、偏置抑制与可复现协议等机制,将“降本”推进为“更稳健、更可控、更可追责”的训练范式。一方面,这些机制增强模型在噪声与分布变化下的表现稳定;另一方面,它们避免将训练阶段的捷径转化为部署阶段的系统性误判风险,从而为医疗、制造与公共治理等高门槛场景中的可信落地提供关键支撑。

3.2 绿色智能

在可信视觉智能的部署过程中,算力与能耗已不再只是效率指标,而逐渐成为系统可用性、稳定性与责任边界的一部分。一方面,大模型训练与推理带来的能耗与碳足迹引发对“可持续AI”的系统性反思,推动社区倡导对能耗与排放进行更透明、可复现的报告与披露[68-69];另一方面,在移动端、边缘侧与在线服务等现实部署中,算力预算、功耗上限、热约束与时延抖动会直接影响输出一致性与服务可靠性。绿色智能因此不宜被狭义理解为“省算力”,更合理的定位是面向可持续落地的部署范式,使更多设备与场景获得稳定可用的智能能力,并在环境与社会层面承担相应技术责任。

从稳健性角度看,资源约束要求模型不仅在理想算力下准确,还需在预算波动、设备降频与并发压力下保持可预期的性能曲线,并呈现可控的退化方式。责任感角度则强调节能应成为可审计、可复现的工程事实,而非难以复现的经验技巧。能耗、延迟与精度的折衷应当可报告,模型在不同硬件与负载下的行为应当可复现、可排查、可治理。换言之,绿色智能的节能不应以不可控漂移为代价,责任也不应脱离可执行的工程边界。

从模型设计路径看,绿色智能首先体现为效率-精度协同的结构化设计。轻量化骨干网络通过硬件感知搜索、算子重构与层级化结构,在降低延迟与功耗的同时维持可用精度,从而减少端侧热降频导致的性能波动并提升服务稳定性。代表性工作包括面向移动端优化的MobileNetV3[70]、兼顾训练效率与参数效率的EfficientNetV2[71],以及通过改进归一化与自监督框架实现同等算力下更强表征的ConvNeXt V2[72]。面向真实设备延迟-精度折衷的混合架构,如文献[73]进一步强调部署可测与性能可落

地的工程稳健性。此类结构化设计将“跑得快”进一步推进为“约束下依然跑得稳”,从而降低设备差异导致的失灵风险,并缓解可信能力仅在高算力群体可得的技术鸿沟。

其次,绿色智能需要借助压缩与近似计算的稳健化机制,将计算节约转化为可控的性能边界。知识蒸馏不仅用于缩小模型规模,还可将大模型的表示能力以更稳定的方式迁移到小模型,缓解小模型在低数据/低算力条件下的方差与不确定性。例如,文献[74]将蒸馏显式融入Transformer训练流程,在保证精度的同时降低训练开销。动态计算策略则提供按需分配算力的路径,如文献[75]的动态Token稀疏化与文献[76]的Token合并,使模型可依据输入难度自适应调整计算量,在不同预算下保持相对稳定的性能曲线,并降低时延抖动。其责任意义在于将“资源不足时系统如何退化”从随机失效转为可设计、可度量、可治理的行为边界,使部署方能够对体验与风险控制给出可验证承诺。

此外,低比特量化是降低推理能耗与显存占用的关键手段,但数值近似可能引入输出漂移与不稳定。近年来相关研究更强调精度损失可界定与误差可校准。文献[77]针对视觉转换器(Vision transformer,ViT)的量化难点进行专门设计,使训练后量化在较低校准成本下实现接近无损的精度保持;文献[78]面向全量化Vision Transformer中LayerNorm与注意力分布导致的量化退化,提出针对性的量化设计以减少精度损失;文献[79]进一步实现覆盖Softmax/高斯误差线性单元(Gaussian error linear unit,GELU)/LayerNorm等非线性算子的纯整数量化推理,使端侧推理路径更可测、更可复现;面向扩散生成模型跨时间步激活分布变化带来的量化失稳,文献[80]通过多时间步校准等机制提升低比特扩散模型的稳定性与可用性。从“服务于人”的角度看,量化不应被视为简单的“牺牲精度换效率”,而应通过校准、鲁棒设计与评测协议将近似误差纳入可解释、可验证的范围,使误差模式保持稳定、可预期、可追溯,避免将节能收益转化为难以解释的用户损失。

综上,绿色智能追求的是资源受限条件下的稳健可信。通过结构设计、蒸馏与动态计算、量化与校准等路径,算力与能耗约束被显式纳入模型能力边界,同时尽可能保证输出一致性、退化可控性与部署可复现性。绿色智能既回应可持续发展的社会责任[68-69],也提升端侧与在线场景的稳定可用性,从而以更低环境与经济成本支撑“服务于人”的可信部署。

3.3 类人图像评价

随着可信视觉智能走向真实应用,“评价”已从离线对比指标演化为连接技术与社会责任的关键枢纽。评价信号决定训练优化方向、生成内容的上线门槛与平台审核标准。若仍主要依赖峰值信噪比(Peak signal-to-noise ratio,PSNR)、结构相似性(Structural similarity index measure,SSIM)等低层误差或结构相似度信号[81],往往难以覆盖人类对自然度与感知相似性的细微差别,易形成“指标高而体验差”的错配,进而引发误导传播、审美操控与不当情绪刺激等责任风险。因此,类人图像评价不应止步于“更像人类打分”,而应成为模型设计服务于人的关键机制,将人类感知与偏好以可量化、可审计、可复现的方式纳入闭环,使输出在质量、语义与价值层面更可控,为内容治理提供可执行的技术抓手。

在责任感层面,类人评价不仅回答“质量如何”,还会影响“是否允许上线、是否需要提示、是否触发复核”。然而,一旦评价模型在数据分布变化、生成失真或跨场景迁移中表现不稳,原本用于把关的责任机制反而可能在关键时刻失灵,引入新的系统性风险。因此,类人图像评价必须同时追求稳健性,在不同内容类型、不同失真形态、不同文化与偏好分布下保持可预期行为,并降低单一指标主导带来的偏置风险。

从技术演进看,类人评价首先体现在感知质量度量由手工规则走向深特征与偏好学习。文献[82]以深特征距离近似感知相似性,显著提升主观一致性;文献[83]统一结构与纹理相似性,在纹理替换与轻微几何不对齐下更稳健。其价值在于使评价信号更贴近真实感知,从而减少模型对不可感知误差的过拟合,并降低指标偏置诱发的不当优化。同时,这类指标也需要在不同内容与轻微分布变化条件下

保持稳定,避免评价漂移削弱上线把关的可信度。

在真实应用中,参考图往往不可获得,使无参考/野外质量评价成为责任场景的重要组成部分,直接影响成像质检、医疗与安防采集有效性以及平台准入阈值。文献[84]通过自适应超网络提升跨内容与跨失真类型泛化能力;文献[85]引入视觉-语言对应关系,将语义背景纳入质量判断,缓解仅凭纹理统计难以解释的主观差异。这些方法推动评价从“给分”走向“可用于决策”,并可进一步支持治理流程,例如在低质或高风险输出上触发降级策略、提示策略或人工复核,而非仅依赖事后纠错。

随着扩散模型等生成式系统成为内容生产基础设施,评价的责任属性被进一步放大,评价不仅影响内容上线,还可能影响推荐与训练数据回流。文献[86]以审美评分分布刻画意见分歧,避免单一分数压缩多样偏好;文献[87]基于用户选择学习可比较的排序评分器,为偏好筛选提供可复现数据基础;文献[88]将专家偏好训练为通用奖励模型,使评价能够作为对齐与治理的训练信号。进一步地,文献[89]将语义、细节与审美等维度拆解建模,支持按维审计与约束,降低奖励黑客与单指标过度优化带来的责任偏差;文献[90]扩展偏好覆盖并推进跨场景稳健评测,为上线评估与迭代提供统一基准。这些工作共同强调两点:评价不仅用于“追求更高分”,更用于形成可治理的门控信号;评价机制若缺乏跨场景稳定性,就难以作为可靠上线标准与训练信号,反而可能放大风险。

面向体验的可信责任还涉及情绪影响的审慎建模。情绪氛围往往决定内容的心理影响与传播效应。文献[91]等“情境情绪”方法将人物与场景上下文共同纳入判断,降低仅凭局部线索进行过度自信推断的风险。在可信框架下,情绪评价更适合作为风险提示与分级治理信号,而非替代价值裁决的自动判定。同时,情绪识别易受文化语境、场景分布与群体差异影响而漂移,因此需要在使用边界、失败提示与审计机制上保持谨慎,避免将不稳定判断固化为平台决策。

综上,类人图像评价通过“感知质量-审美偏好-情绪体验”的层次化度量,将人类体验以可计算形式嵌入视觉智能闭环。一方面,它为训练、筛选与部署提供可追溯的证据链;另一方面,它要求把关信号在分布变化与生成失真条件下仍可靠执行。当评价从单纯“指标”升级为“门控与治理机制”时,模型设计才能更充分地体现“服务于人”,使系统输出更符合体验预期,并在风险与争议面前保持可控与可问责。

4 系统应用受控于人

继前章探讨了如何在认知空间内构建稳健且负责的模型设计之后,本节将视野进一步从逻辑中枢转向物理空间的落地实现。系统应用受控于人、立足于以人为中心的智能系统设计理念,将控制作为贯穿系统设计、部署与运行过程的系统性保障机制,其核心目标在于确保人类在视觉智能系统自主性不断增强的过程中,始终处于认知与决策的反馈闭环中心,推动认知空间透明性与物理空间安全性的协同提升。

这一理念在价值取向上,深度呼应了斯坦福大学李飞飞教授所倡导的“人工智能应当增强人性而非削弱判断”的主张[92]。她强调,智能系统的目标不应是单纯的技术替代,而应通过更接近人类认知逻辑的环境理解与空间建模,弥合人机间的认知鸿沟,从而从根本上增强人类对系统行为的预判与决策能力。与此同时,该理念也契合了卡内基梅隆大学赵鼎教授针对开放世界长尾场景提出的干预机制[93]。他指出,在复杂且不可完全建模的真实环境中,视觉系统的安全性并非源于单次算法精度的提升,而在于系统是否具备清晰的能力边界评估以及面向人类的可干预设计,以确保在极端失效模式下人类依然保有最终的修正与控制权。相关研究共同指向一个核心结论,即系统能力的提升必须与人类理解及控制能力的增强同步对齐,否则高性能反而可能放大潜在的失控风险[94]。近年来,随着自动驾驶、机器人控制等高自主系统的规模化应用,受控于人已从价值层面的理念倡议转变为由实际部署驱

动的核心工程议题[95-96]。这类系统运行在高度不确定的物理环境中,其失效往往关联重大社会风险,因此仅依赖事后解释难以支撑对复杂系统的实时控制[97-98],须通过显式、可操作的交互机制,提升人类对系统状态、风险边界与失效条件的实时理解与调节能力。

基于上述需求,本节将透明性与安全性作为系统应用的两条相互嵌套的主线,从3个维度构建“受控于人”的技术框架,如图5所示。首先,人机可靠感知侧重于通过可解释视觉分析将机器的判断逻辑转化为人类可理解、可验证的信息,旨在消除信息不对称并奠定“人懂机器”的信任基线。与此同时,人机智能交互发挥信息枢纽作用,借助视觉引导的多模态对话与反馈机制建立双向语义通道,使系统逻辑能够深度对齐人类意图并推动“机器懂人”的工程落地。此外,人机混合增强重点在于实现系统对人的状态与意图的持续建模,通过将人类的高层目标与安全约束显式注入执行回路,并结合不确定性门控机制确保机器在关键场景下安全让渡控制权,从而达成“人机共生”的协同形态。这3个维度共同构成了一个以人类理解、监督与最终决策能力为中心的统一问题框架,在保障视觉智能技术在物理世界中稳态运行的同时,更能在人类意图的指引下实现安全、可持续的技术落地。

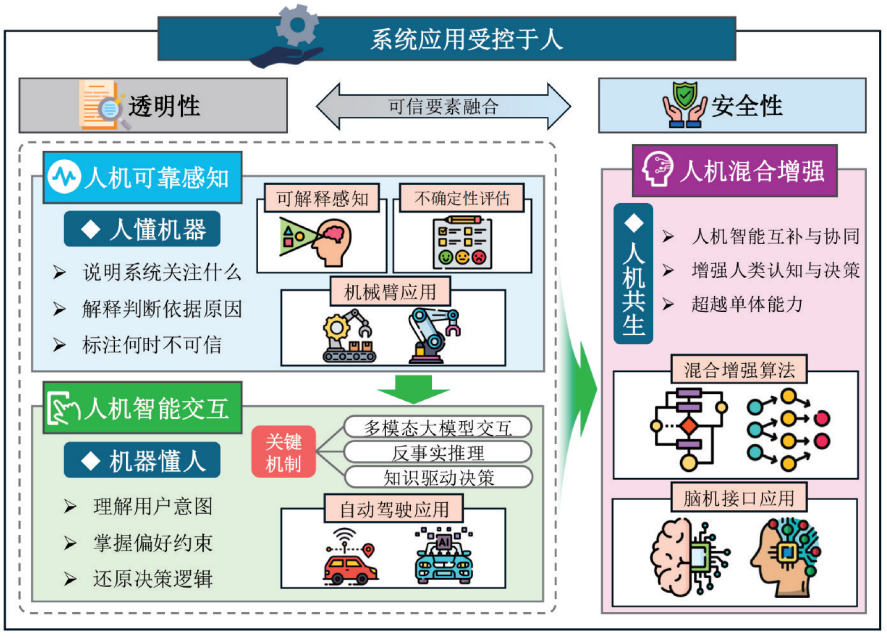


图5 系统应用受控于人

Fig.5 Human-controlled system deployment

4.1 人机可靠感知

人机可靠感知的核心在于“增强理解,赢得信任”,是确保视觉智能技术受控于人的基础环节。在具身智能系统中,感知不再是孤立的语义标签,而是直接驱动机械臂抓取、车辆避障等物理行为的指令源。由于具身智能体运行在不可完全建模的物理世界中,感知一旦失效,往往会通过执行层被迅速放大,导致碰撞、损毁等不可逆后果[99]。正因如此,系统不仅需要输出传统意义上的感知结果,还必须提供与人类认知相对齐的证据,即确保机器对环境的理解方式与人类直觉和安全预期保持一致。这种从追求性能最优向追求可判信的范式演进,是实现人类在环监督的前提,也是高自主系统得以安全部署的必要条件[100]。

从技术视角看,这一以可信性为导向的人机可靠感知目标具体可分解为两个相互补充的维度。其

一,系统是否能够向人类解释为何做出当前感知判断,即认知的透明性;其二,在复杂、遮挡或长尾场景中,系统是否能够明确探测其能力边界与安全状态,即行为的安全性。前者对应可解释视觉分析,通过逻辑回溯与语义对齐消除人机间的认知黑箱。后者则对应安全边界探测与风险评估,通过将感知的局限性转化为显式的风险指标,确保系统在极端或长尾场景下具备自我暴露风险的能力。这两条路径共同构成了人机可靠感知向透明性、安全性对齐的核心实现路径。

在透明性维度,早期的可解释视觉研究主要侧重于缓解深度模型的“黑箱问题”,通过后验可视化手段将模型内部关注区域展示给人类。例如,文献[101]等方法利用梯度信息生成显著性图,为人类提供“模型在看哪里”的直观线索。这类方法在一定程度上提升了模型透明度,但其解释仍停留在像素或局部区域的重要性层面,难以直接对应具身任务中涉及的对象、关系与物理语义。为缩小这一认知鸿沟,研究逐步转向结构化与语义化解释。概念激活向量[102]通过引入人类可理解的高层语义概念,量化特定语义因素对模型决策的影响,使解释对象从像素贡献提升为概念作用。原型驱动的可解释模型则通过在模型结构中显式嵌入可解释原型,使决策过程能够以“与哪些具体实例相似”的方式呈现,从而更贴近人类的类比式认知。

以具身智能机械臂操作领域相关工作为例,文献[103]通过将预测结果与一组语义一致的视觉原型进行匹配,使模型判断依据可观测。这类基于原型的结构化解释方法已被系统性总结为当前神经网络可解释性的重要研究方向之一[104]。进一步地,因果推理框架的引入标志着可解释性从可视化层面迈向可靠性层面的关键突破。通过反事实分析与干预建模,相关方法能够识别真正驱动系统决策的物理因果因素,并将其与由数据分布偏差诱发的虚假相关性区分开来,从而为具身智能系统在复杂、动态环境中的感知判断提供更稳健的认知基础与安全保障[105]。随着具身大模型的发展,可解释性开始从事后解释演进为面向任务的认知对齐机制[106]。文献[107]提出了一种认知对齐的视觉、语言、动作和框架,通过指令驱动的多阶段路由机制,在结构层面将人类任务意图显式注入感知与决策过程;文献[108]针对现有多模态大模型在长时序机器人操纵任务中缺乏规划、可供性感知与轨迹预测能力的问题,结合多阶段训练策略,使模型能够将抽象操纵指令逐步映射为具备可执行意义的中间表示。这种从高层意图到具体操作要素的逐级展开,使复杂操纵决策过程不再完全隐含于黑箱推理中,而具备可理解的中间结构。文献[109]通过视觉指令预训练以视觉目标替代文本描述来指定操纵任务,并引入稀疏点流预测作为中间表征,从而减少语言指令在复杂操纵场景中的歧义。这种以视觉中间状态明确任务目标的方式,使系统正在对齐的目标与预期演化方向对人类更加直观可见,为人类理解与监督具身决策过程提供了更符合多通道认知方式的感知依据。

仅回答系统为何如此判断,仍不足以支撑自动驾驶与机器人操纵等高风险系统的安全运行。在真实物理环境中,人类同样关心系统在何种条件下可能失效、是否已接近能力边界。安全性风险评估正是回应这一问题的关键技术,它通过将模型内部的认知局限显式化,使潜在风险在转化为物理行为之前对人类可见。现有研究通常将感知风险区分为知识局限引发的内部风险与环境噪声、遮挡引发的外部风险,前者理论上可通过数据扩充缓解,后者则具有不可消除性[110]。针对这些风险,研究者逐步发展出一系列安全性量化方法,用以在系统层面刻画其行为边界。从方法上看,现有的安全性建模技术大体可归纳为5类:其一是以贝叶斯神经网络为代表的概率建模方法,通过推断权重的后验分布实现对感知可靠性的量化评估[111-112];其二是深度集成方法,通过多模型预测的一致性分析量化感知判断的潜在冲突,从而提升系统对极端场景的安全性识别能力[113-114];此外,还有推理期采样、后处理风险学习以及通过生成离群样本显式探测错误边界的数据增强方法(如OpenMix[115]),这些共同构成了较为完整的安全性评价体系。这些方法的共同价值并不在于进一步提升单点预测精度,而在于为系统输出附加可操作的风险边界,使人类能够判断当前感知结果是否仍处于安全可信区间,从而为是否介入、降级或中

止系统行为提供明确依据。

在具身智能应用领域中,安全性边界的显式建模决定了系统是否具备自我暴露风险的能力。以机器人操纵为例,文献[116]通过将强化学习信号与视觉-语言操纵策略深度耦合,增强了系统在物体属性变化下的稳定性,降低了感知偏差演化为任务失败的风险。文献[117]则通过引入显式的空间记忆机制,有效缓解了动态遮挡引发的瞬时感知波动,提升了感知与动作在时间维度上的可靠性与可预测性。文献[118]提出的双系统结构,在物理动作生成之前显式进行视觉潜空间规划,使系统在执行前能够暴露其规划路径与推理一致性,为人类监督者提供了审视与评估的机会。当不同候选规划之间出现明显分歧时,系统能够通过延迟执行与自我修正机制避免不成熟决策直接转化为不可逆的物理动作。文献[119]通过对象中心的轨迹表征,将复杂策略解耦为人类直觉可判断的姿态轨迹,使操作者能够直接评估操作的合理性与安全性。而文献[120-121]对通用动作空间的规范,进一步缩小了感知、规划与执行之间的认知断层,保障了系统整体的安全性。

综上所述,这一系列进展表明,视觉智能技术正在从追求绝对性能转向追求可用、可信与可控的演进。通过将感知结果、解释依据与风险边界整合为完整的安全证据链,具身智能系统即使在复杂长尾场景下,也能够持续处于人类的理解与评估之下,为其大规模、长期可靠部署奠定了坚实基础。从“系统应用受控于人”的视角看,这一转变的关键在于将感知模块由黑箱组件转化为认知接口。可解释视觉分析回答系统为何如此判断,安全性风险评估进一步揭示了在何种条件下不应信任该判断。二者在工程层面的融合,使机器的输出被转译为人类可理解、可操作的可判性信息,从而构建起透明、可验证且可控的人机可靠感知闭环。这一感知闭环的建立初步解决了“人懂机器”的信任难题,为人类行使监督与决策权提供了必要的信息支撑。

4.2 人机智能交互

在人机可靠感知建立的“人懂机器”信任基线之上,人机智能交互进一步聚焦于如何实现“机器懂人”,旨在构建人机双向的语义对齐空间。人机交互(Human-computer interaction,HCI)是指人与计算机之间通过某种对话语言和交互方式,为完成特定任务而进行的信息交换过程[122]。它不仅关注传统的输入输出设备设计,更强调系统如何理解用户需求并以直观高效的方式反馈结果。随着智能技术的发展,HCI 逐渐演变为人机智能交互,要求机器具备洞悉需求的能力,通过视觉智能、多模态融合等技术,实现从被动响应到主动服务的转变。这种转变不仅提升了系统的认知透明性,更通过精准的意图识别保障了物理操作的安全性。

从历史演进看,HCI 的发展轨迹本质上是认知鸿沟不断缩减、系统透明度持续提升的过程。早期以穿孔卡片、纸带与控制面板开关为主的交互方式[123],完全依赖专业人员的机械操作,交互逻辑极为原始且封闭。1959 年 Shackel[124]关于控制台人机工程学的研究与 1960 年 Licklider 等[125]提出的“人机共生”启蒙思想,开启了对人机深度协作的探索。1969 年第一届人机系统国际大会的举办与《国际人机研究》的创刊,标志着该领域正式进入学科化进程[126]。从命令行界面到 20 世纪 80 年代图形用户界面的引入[127],交互的直观性极大降低了使用门槛。进入 21 世纪,随着语音识别[128]、手势控制[129]、脑机接口[130]以及增强现实与虚拟现实[131]等技术的应用,自然交互成为主流。在这种演变中,隐形界面与情感计算使系统能够感知人类的情绪与状态,从而在透明化、智能化的基础上,为实现个性化且安全的行为反馈提供了可能。

这种从“人类理解机器”向“机器理解人类”的逻辑转变,在自动驾驶领域表现得尤为突出。早期的交互逻辑以功能菜单为中心,增加了用户的认知负担。随着大语言模型(Large language model, LLM)的引入,交互进入了语义对齐阶段。文献[132]通过将视觉特征注入多模态大模型,使系统能以自然语言陈述感知结论,实现了从人类解读数据向机器主动解释的跨越,显著增强了决策过程的透明性。在

此基础上,交互进一步扩展到知识驱动的推理层面,文献[133]通过显式引入推理与反思模块,使系统能够进行因果分析并总结经验,使人类能够通过审查反思信息行使高层监督权。为了确保复杂场景下的安全性,交互技术开始引入更深层的逻辑推演与干预机制。文献[134]引入反事实推理机制,使系统能回答假设性问题,将风险推演过程透明化。同时,文献[135]对视觉-语言模型的交互能力进行细粒度评测,确保其输出符合人类交通规则与安全常识。轻量化架构VLDrive[136]则确保在受限算力下仍能解释决策依据,保障了人类在回路中的核心地位。进一步地,文献[137]提出的基于快慢系统的自适应协同机制,在复杂长尾场景下通过引入人类介入,从机制上保留了对关键决策的最终控制权。

总之,HCI的历程是从工具性协作迈向智能共生的演进。从“受控于人”的视角看,这些交互技术共同推动了人类角色从执行者向监督决策者的转变,确保系统始终服从于人的意图、价值与安全边界。虽然HCI面临隐私保护、数据安全等新挑战,但其核心逻辑始终围绕着提升系统的透明度与安全性。未来,人机智能交互将继续朝着自然化与人性化迈进,将视觉智能与人类智慧进行深度融合。

4.3 人机混合增强

继人机可靠感知确立了“人懂机器”的信任基准,以及人机智能交互实现了“机器懂人”的语义对齐之后,视觉智能系统迈向了受控于人的“人机共生”阶段,即人机混合增强。感知与交互解决了人机之间的信息对称与意图对齐问题,而混合增强则聚焦于物理空间的行为共生,旨在通过人机权责的动态分配与智能耦合,构建一个执行层面的受控闭环。

人机混合增强智能的核心思想是在智能机器中嵌入类人的认知能力或角色。当人与机器协作时,二者智能结合,从而增强人类的智力与认知能力,实现人或机单独无法完成的目标[138-139]。自2017年中国将混合增强智能纳入《新一代人工智能发展规划》以来,该领域已在协作感知、计算前移与环境自适应等基础理论层面实现了深度布局[140],并广泛渗透至教育等数字化转型范式之中。随着大语言模型的爆发,混合增强不再是简单的能力叠加,而是机器计算逻辑与人类涵盖感知、注意、记忆、推理和决策在内的完整认知过程的深度架构对齐[141]。

这种协同架构的有效运行,依赖于透明性驱动的认知增强与安全性支撑的决策门控。在认知维度,机器需从被动工具演进为具备意图理解能力的团队成员,通过引入记忆、反思与规划机制模拟人类认知过程,为构建逻辑透明的协同代理提供了可行范式。Sanfilippo等[142]系统梳理了人机系统从人机交互向人机协作(Human-machine collaboration, HMC)及人机团队的演进过程,指出机器应从被动工具转变为具备意图理解与联合任务能力的团队成员,为HMC提供了结构化框架。在认知层增强方面,Park等[143]提出的生成式代理通过引入记忆、反思与规划机制,使基于大语言模型的智能体能够模拟人类认知过程,表现出稳定的个体行为与群体涌现特性,为基于认知计算的混合增强智能(Cognitive computing based hybrid-augmented intelligence, CC-HAI)中“认知队友”的构建提供了可行范式。而在安全性维度,基于人类在环的混合增强智能(Human-in-the-loop based hybrid-augmented intelligence, HITL-HAI)强调了人类判断的最终主导作用。通过将模型置信度与人类信心进行对齐,系统能够显著提升决策的可发现性与信任校准能力[144]。Corvelo等[145]从理论上证明,仅依赖模型置信度难以支持人类形成最优决策策略,而通过将模型输出与人类自身信心对齐,可显著提升决策可发现性与信任校准能力。在此基础上,文献[146]提出的双范式混合增强智能(Dual-paradigm hybrid-augmented intelligence, DP-HAI)框架系统融合CC-HAI与HITL-HAI,通过结构化提示将大语言模型转化为领域角色,并结合Z数增强的云建模方法对不确定性进行表征,实现人机团队组合与优化的统一建模与验证。进一步地,在团队层协同支持方面,Endsley[147]提出的态势感知导向设计(Situation awareness oriented design, SAOD)方法从任务态势、代理态势和团队态势3个层面,强调透明性与可解释性在支持人机团队共享态势感知与信任校准中的核心作用。

在具体应用层面,植入式脑机接口(Implanted brain-computer interface, iBCI)为混合增强智能提供了一种将内隐认知状态直接嵌入系统控制回路的受控范式。作为一种典型的人机闭环系统,iBCI架构主要由用于检测大脑信号的电极、负责信号预处理与解码的计算单元、假体或外部应用设备以及反馈回路等4个核心组件构成[148]。近年来的临床研究回顾表明,该类系统已能够在长期稳定运行条件下支持高精度运动控制、自然语言解码以及感觉-运动双向闭环,从而在执行层面实现人与机器的深度协同与责任对齐[148]。自1998年完成首例人类长期植入并实现游标控制以来[149],以BrainGate研究组为代表的多个团队在长达二十余年的研究与随访中,系统验证了皮层内接口在驱动游标、操作多关节机器人臂以及执行日常电子设备指令等任务中的长期稳定性与可靠性[150-151]。为进一步增强系统运行过程中的透明性与人类主导感,近年来的研究引入了感觉反馈机制,通过皮层内微刺激恢复触觉感知,构建起感觉-运动双向闭环控制逻辑[152]。在行为恢复层面,通过将神经信号直接转化为功能性电刺激指令,研究者已成功实现了瘫痪肢体运动的协同还原[153],并进一步通过脑-脊髓接口实现了步行能力的自然重建[154]。与此同时,结合大语言模型背景下的最新言语解码研究,iBCI在交流效率方面取得突破性进展,其文本或语音输出速度已提升至每分钟62~79词的水平,显著增强了人机之间的语义对齐能力[144,155-156]。从“受控于人”的视角看,iBCI系统的安全性深度依赖于对其运行风险的持续监测与评估。相关中期安全性研究表明,系统能够在长期运行过程中识别潜在失效模式并动态校准信任边界,从而确保共享控制始终处于人类能力与责任边界的约束之下[157]。这种以内隐认知为核心的人机协同控制范式,使得系统在冲突情形或认知能力受限时能够主动让渡控制权,标志着iBCI已从早期功能验证阶段迈向以长期稳定性、安全性与人类主导为核心的受控混合增强智能系统。

综上所述,人机混合增强不仅实现了人机在任务层面的协同,更在执行层面构建了权责明确的受控闭环。在这种模式下,透明性保障了人类对机器推理过程的实时监督,而基于状态感知的安全性门控则确立了人类在不确定场景下的最终决策主权。通过将人类的认知主导性与机器的计算辅助性有机结合,混合增强智能确保了视觉系统在追求极致性能的同时,始终被约束在安全可控的行为边界之内。这不仅是技术向智能共生演进的必然路径,更是视觉智能最终回归于人类福祉、实现价值对齐的工程保障。

5 结束语

本文围绕“以人为中心的可信视觉智能”展开综述与框架化总结,面向视觉系统从感知走向自主决策与物理执行的新阶段,系统梳理了隐私、公平、稳健、透明与安全等可信风险在视觉全生命周期中的表现与应对策略。针对现有综述多从单一可信属性出发进行并列式梳理、缺乏维度耦合分析,且对“人”在不同技术层级中的角色演化与动态约束刻画不足、难以解释视觉感知-决策-物理行为链路中的风险传导机理等问题,本文从计算机视觉视角重构知识组织方式,提出以信息空间-认知空间-物理空间为主线的统一框架,并以“关注于人-服务于人-受控于人”为递进逻辑,建立可信要素、系统层级与人类角色之间的映射关系,为数据分析、模型设计与系统应用提供结构统一且可执行的研究路线图。本文一方面将分散的可信要素纳入同一框架下进行全链路组织,强调可信机制在数据、模型与系统部署阶段的协同设计;另一方面突出视觉智能的领域特性,即感知输出往往直接触发决策与物理执行,从而使模型失效可能经由人的误判被放大为现实伤害,进而为高影响场景的可信评估与工程落地提供更贴合应用的分析视角。与既有工作相比,本文不再停留在“按属性分章节”的静态综述结构,而是以“人”的多重角色为核心主线,面向系统全生命周期给出跨维度、跨层级的统一组织与解释框架。

以人为中心的可信视觉智能研究不仅关注人工智能的技术本身,更强调如何将其与人类需求紧密结合。在这一背景下,提出以下展望:

(1)常识驱动将成为未来研究的重要方向。强调符合人类常识不仅是为了提升人工智能系统的效

率,更是为了促使网络主动辨别数据的真假与优劣。通过引入人类的常识,人工智能可以更有效地学习并发现可信数据,从而确保技术的合理应用,减少潜在的误导性和偏差。

(2)关注于人是实现可信视觉智能的关键。需要考虑多模态、多视角和多粒度的信息融合,充分利用不同来源的数据,通过层次分析来确保所产生的内容可信且符合人类认知。这样的方法不仅提升了信息处理的准确性,还能使系统在复杂环境下更好地理解和响应人类的需求。

(3)服务于人是可信视觉智能的追求目标。研究应致力于实现人类视觉感知与智能系统之间的主客观一致性。构建可解释性强的可信模型,使得人工智能系统能够真实地服务于人的需求,增强用户对系统的信任感和依赖性。

(4)受控于人是确保人工智能系统安全性和可靠性的基础。应重视公平性与隐私保护,确保构建的智能系统能够在人的控制之下,具备必要的安全性和可靠性,避免潜在的风险和误用。通过制定有效的规范与标准,使技术的发展始终沿着有利于人类社会的方向前进。

未来的研究应继续围绕这些核心思想展开,确保技术的进步不仅是科学的突破,更是对人类社会的积极贡献。只有在以人为中心的框架下,可信视觉智能才能真正实现其价值,促进社会的可持续发展与繁荣。

参考文献:

[1] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]//Proceedings of International Conference on Neural Information Processing Systems (NIPS). Stateline, USA: MIT Press, 2012: 1097-1105.

[2] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Proceedings of International Conference on Neural Information Processing Systems (NIPS). Montreal, Canada: MIT Press, 2014: 2672-2680.

[3] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of International Conference on Neural Information Processing Systems (NIPS). California, USA: MIT Press, 2017: 6000-6010.

[4] LI J, LI D, SAVARESE S, et al. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models[C]//Proceedings of International Conference on Machine Learning (ICML). Honolulu, HI, USA: [s.n.], 2023: 19730-19742.

[5] CARON M, TOUVRON H, MISRA I, et al. Emerging properties in self-supervised vision transformers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, South Korea: IEEE, 2021: 9650-9660.

[6] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[C]//Proceedings of International Conference on Neural Information Processing Systems (NIPS). [S.l.]: MIT Press, 2020: 6840-6851.

[7] WOLF V, MAIER C. ChatGPT usage in everyday life: A motivation-theoretic mixed-methods study[J]. International Journal of Information Management, 2024, 79: 102821.

[8] SCHNEPF J, ENGIN T, ANDERER S, et al. Studies on the use of Large language models for the automation of business processes in enterprise resource planning systems[C]//Proceedings of International Conference on Applications of Natural Language to Information Systems. Berlin, Germany: Springer, 2024: 16-31.

[9] ZHAO X, WANG L, ZHANG Y, et al. A review of convolutional neural networks in computer vision[J]. Artificial Intelligence Review, 2024, 57(4): 57-99.

[10] ZALA K, THUMAR D, THAKKAR H K, et al. A survey and identification of generative adversarial network technology-based architectural variants and applications in computer vision[J]. International Journal of System Assurance Engineering and Management, 2024, 15(9): 4594-4615.

[11] TUCUDEAN G, BUCOS M, DRAGULESCU B, et al. Natural language processing with transformers: A review[J]. PeerJ Computer Science, 2024, 10: e2222.

[12] AWAIS M, NASEER M, KHAN S, et al. Foundation models defining a new era in vision: A survey and outlook[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(4): 2245-2264.

[13] MENG C, GRIESEMER S, CAO D, et al. When physics meets machine learning: A survey of physics-informed machine learning[J]. Machine Learning for Computational Science and Engineering, 2025, 20(1): 20.

[14] TAN Z, LI D, WANG S, et al. Large language models for data annotation and synthesis: A survey[C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP). Florida, USA: ACL, 2024: 930-957.

[15] LI Z, YU Z, LAN S, et al. Is ego status all you need for open-loop end-to-end autonomous driving?[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2024: 14864-14873.

[16] PAPAGIANNIDIS E, MIKALEF P, CONBOY K. Responsible artificial intelligence governance: A review and research framework[J]. The Journal of Strategic Information Systems, 2025, 34(2): 101885.

[17] GRIFFEN Z, OWENS K. From "human in the loop" to a participatory system of governance for AI in healthcare[J]. The American Journal of Bioethics, 2024, 24(9): 81-83.

[18] WANG Z, LI X, ZHU H, et al. Revisiting adversarial training at scale[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2024: 24675-24685.

[19] TRAMÈR F, KAMATH G, CARLINI N. Position: Considerations for differentially private learning with large-scale public pretraining[C]//Proceedings of International Conference on Machine Learning (ICML). Vienna, Austria: ACM, 2024: 48453-48467.

[20] HONG M, YUN J, JEON I, et al. FedAvP: Augment local data via shared policy in federated learning[C]//Proceedings of International Conference on Neural Information Processing Systems (NIPS). Vancouver, Canada: MIT Press, 2024: 18090-18121.

[21] PAREKH J, KHAYATAN P, SHUKOR M, et al. A concept-based explainability framework for large multimodal models [C]//Proceedings of International Conference on Neural Information Processing Systems (NIPS). Vancouver, Canada: MIT Press, 2024: 135783-135818.

[22] KOWALD D, SCHER S, PAMMER-SCHINDLER V, et al. Establishing and evaluating trustworthy AI: Overview and research challenges[J]. Frontiers in Big Data, 2024, 7: 1467222.

[23] 安永(中国)企业咨询有限公司, 上海市人工智能与社会发展研究会. 可信人工智能治理白皮书[EB/OL]. (2025-01-13)[2025-12-30]. https://pdf.dfcfw.com/pdf/H3_AP202501141641916724_1.pdf.
Ernst & Young (China) Enterprise Consulting Co., Ltd., Shanghai association for artificial intelligence and social development studies. White paper on governance of trustworthy artificial intelligence[EB/OL]. (2025-01-13)[2025-12-30]. https://pdf.dfcfw.com/pdf/H3_AP202501141641916724_1.pdf.

[24] LIU H, WANG Y, FAN W, et al. Trustworthy AI: A computational perspective[J]. ACM Transactions on Intelligent Systems and Technology, 2022, 14(1): 1-59.

[25] MCCORMACK L, BENDECHACHE M. A comprehensive survey and classification of evaluation criteria for trustworthy artificial intelligence[J]. AI and Ethics, 2025, 5(3): 1973-1994.

[26] WEI W, LIU L. Trustworthy distributed AI systems: Robustness, privacy, and governance[J]. ACM Computing Surveys, 2025, 57(6): 1-42.

[27] MA J, WANG T, LIU M, et al. DCHM: Depth-consistent human modeling for multiview detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Hawaii, USA: IEEE, 2025: 7731-7740.

[28] YAMANE T, MASUMURA R, SUZUKI S, et al. MVTrajecter: Multi-view pedestrian tracking with trajectory motion cost and trajectory appearance cost[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Hawaii, USA: IEEE, 2025: 13270-13280.

[29] KIM T, SHIN S, YU Y, et al. Causal mode multiplexer: A novel framework for unbiased multispectral pedestrian detection [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington, USA: IEEE, 2024: 26784-26793.

[30] ZHANG Y, ZENG W, JIN S, et al. When pedestrian detection meets multi-modal learning: Generalist model and benchmark dataset[C]//Proceedings of European Conference on Computer Vision (ECCV). Milan, Italy: Springer, 2024: 430-448.

[31] CHEN Y, GUO J, GUO S, et al. Neuron: Learning context-aware evolving representations for zero-shot skeleton action

recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Conference (CVPR). Tennessee, USA: IEEE, 2025: 8721-8730.

[32] WANG H, MA X, KUANG J, et al. Heterogeneous skeleton-based action representation learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Conference (CVPR). Tennessee, USA: IEEE, 2025: 19154-19164.

[33] ZHU M, HO ESL, CHEN S, et al. Geometric features enhanced human-object interaction detection[J]. IEEE Transactions on Instrumentation and Measurement, 2024, 73: 1-14.

[34] BEN-SHABAT Y, SHROUT O, GOULD S. 3DInAction: Understanding human actions in 3D point clouds[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington, USA: IEEE, 2024: 19978-19987.

[35] KIM M, YE D, SU Y, et al. SapiensID: Foundation for human recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Conference (CVPR). Tennessee, USA: IEEE, 2025: 13937-13947.

[36] ZHOU K, YANG F, WANG S, et al. M-SpecGene: Generalized foundation model for RGBT multispectral vision[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Hawaii, USA: IEEE, 2025: 7861-7872.

[37] YAO H, YANG B, HUANG W, et al. Unsupervised visible-infrared person re-identification under unpaired settings[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Hawaii, USA: IEEE, 2025: 11916-11926.

[38] YUAN C, ZHANG G, MA C, et al. From poses to identity: Training-free person re-identification via feature centralization [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Conference (CVPR). Tennessee, USA: IEEE, 2025: 24409-24418.

[39] SELVARAJU P, ABREVAYA V F, BOLKART T, et al. OFER: Occluded face expression reconstruction[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Conference (CVPR). Tennessee, USA: IEEE, 2025: 26985-26995.

[40] YANG X, TAKETOMI T, ENDO Y, et al. FreeUV: Ground-truth-free realistic facial UV texture recovery via cross-assembly inference strategy[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Conference (CVPR). Tennessee, USA: IEEE, 2025: 326-337.

[41] LYU W, ZHOU Y, YANG M H, et al. FaceLift: Learning generalizable single image 3D face reconstruction from synthetic heads[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Hawaii, USA: IEEE, 2025: 12691-12701.

[42] LI H, FENG Y, XUE S, et al. UV-IDM: Identity-conditioned latent diffusion model for face UV-Texture generation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington, USA: IEEE, 2024: 10585-10595.

[43] TOMAŠEVIĆ D, BOUTROS F, LIN C, et al. ID-Booth: Identity-consistent face generation with diffusion models[C]//Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG). Florida, USA: IEEE, 2025: 1-10.

[44] WANG H, MAI X, TAO Z, et al. D2SP: Dynamic dual-stage purification framework for dual noise mitigation in vision-based affective recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Conference (CVPR). Tennessee, USA: IEEE, 2025: 19218-19229.

[45] TAN Y, XIA H, SONG S. Robust consistency learning for facial expression recognition under label noise[J]. The Visual Computer, 2025, 41(4): 2655-2667.

[46] ZHANG B, WANG X, WANG C, et al. Dynamic stereotype theory induced micro-expression recognition with oriented deformation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Conference (CVPR). Tennessee, USA: IEEE, 2025: 10701-10711.

[47] ZHAO L, XUAN J, LOU J, et al. Context-aware academic emotion dataset and benchmark[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Hawaii, USA: IEEE, 2025: 13859-13868.

[48] LIU D, WANG Z, PENG C, et al. Thinking racial bias in fair forgery detection: Models, datasets and evaluations[C]//

Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Pennsylvania, USA: AAAI, 2025: 5379-5387.

[49] KASHIANI H, TALEMI N A, AFGHAH F. FreqDebias: Towards generalizable deepfake detection via consistency-driven frequency debiasing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Tennessee, USA: IEEE, 2025: 8775-8785.

[50] LIN L, HE X, JU Y, et al. Preserving fairness generalization in deepfake detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington, USA: IEEE, 2024: 16815-16825.

[51] LIN X, WANG S, CAI R, et al. Suppress and rebalance: Towards generalized multi-modal face anti-spoofing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington, USA: IEEE, 2024: 211-221.

[52] YANG J, LIN X, YU Z, et al. Dadm: Dual alignment of domain and modality for face anti-spoofing[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Hawaii, USA: IEEE, 2025: 12045-12056.

[53] CAI Y, LI J, LI Z, et al. DeepShield: Fortifying deepfake video detection with local and global forgery analysis[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Hawaii, USA: IEEE, 2025: 12524-12534.

[54] SUN K, CHEN S, YAO T, et al. Towards general visual-linguistic face forgery detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Tennessee, USA: IEEE, 2025: 19576-19586.

[55] LI Z, ZHAO T, XU X, et al. Optimal transport-guided source-free adaptation for face anti-spoofing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Tennessee, USA: IEEE, 2025: 24351-24363.

[56] GE X, LIU X, YU Z, et al. Difffas: Face anti-spoofing via generative diffusion models[C]//Proceedings of European Conference on Computer Vision (ECCV). Milan, Italy: Springer, 2024: 144-161.

[57] SHAO R, PERERA P, YUEN P C, et al. Federated generalized face presentation attack detection[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(1): 103-116.

[58] SOHN K, BERTHELOT D, LI C L, et al. FixMatch: Simplifying semi-supervised learning with consistency and confidence [C]//Proceedings of International Conference on Neural Information Processing Systems (NIPS). [S.l.]: MIT Press, 2020: 596-608.

[59] ZHANG B, WANG Y, HOU W, et al. FlexMatch: Boosting semi-supervised learning with curriculum pseudo labeling[C]//Proceedings of International Conference on Neural Information Processing Systems (NIPS). [S.l.]: MIT Press, 2021: 18408-18419.

[60] WANG Y, CHEN H, HENG Q, et al. FreeMatch: Self-adaptive thresholding for semi-supervised learning[C]//Proceedings of International Conference on Learning Representations (ICLR). Kigali, Rwanda: ICLR, 2023.

[61] CHEN H, TAO R, FAN Y, et al. SoftMatch: Addressing the quantity-quality trade-off in semi-supervised learning[C]//Proceedings of International Conference on Learning Representations (ICLR). Kigali, Rwanda: [s.n.], 2023.

[62] LIU Y C, MA C Y, HE Z, et al. Unbiased teacher for semi-supervised object detection[C]//Proceedings of the International Conference on Learning Representations (ICLR). Virtual Event: [s.n.], 2021.

[63] XU M, ZHANG Z, HU H, et al. End-to-end semi-supervised object detection with soft teacher[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). [S.l.]: IEEE, 2021: 3060-3069.

[64] LV J, LIU B, FENG L, et al. On the robustness of average losses for partial-label learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(5): 2569-2583.

[65] WANG W, WU D D, WANG J, et al. Realistic evaluation of deep partial-label learning algorithms[C]//Proceedings of the 13th International Conference on Learning Representations (ICLR). Singapore, Singapore: [s.n.], 2025.

[66] GU X, LIN T Y, KUO W, et al. Open-vocabulary object detection via vision and language knowledge distillation[C]//Proceedings of International Conference on Learning Representations (ICLR). Virtual Event: [s.n.], 2022.

[67] ZHOU X, GIRDHAR R, JOULIN A, et al. Detecting twenty-thousand classes using image-level supervision[C]//Proceedings of European Conference on Computer Vision (ECCV). Tel Aviv, Israel: Springer, 2022: 350-368.

[68] STRUBELL E, GANESH A, MCCALLUM A. Energy and policy considerations for deep learning in NLP[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational

Linguistics, 2019: 3645-3650.

[69] DE VRIES A. The growing energy footprint of artificial intelligence[J]. Joule, 2023, 7(10): 2191-2194.

[70] HOWARD A, SANDLER M, CHEN B, et al. Searching for mobileNetV3[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, South Korea: IEEE, 2019: 1314-1324.

[71] TAN M, LE Q. EfficientNetV2: Smaller models and faster training[C]//Proceedings of the 38th International Conference on Machine Learning (ICML). Virtual Event: [s.n.], 2021: 10096-10106.

[72] WOO S, DEBANTH S, HU R, et al. Convnext v2: Co-designing and scaling convnets with masked autoencoders[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023: 16133-16142.

[73] VASU P K A, GABRIEL J, ZHU J, et al. Fastvit: A fast hybrid vision transformer using structural reparameterization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023: 5785-5795.

[74] TOUVRON H, CORD M, DOUZE M, et al. Training data-efficient image transformers & distillation through attention[C]//Proceedings of the 38th International Conference on Machine Learning (ICML). [S.l.]: ACM, 2021: 10347-10357.

[75] RAO Y, ZHAO W, LIU B, et al. DynamicViT: Efficient vision transformers with dynamic token sparsification[C]//Proceedings of International Conference on Neural Information Processing Systems (NIPS). [S.l.]: MIT Press, 2021: 13937-13949.

[76] BOLYA D, FU C Y, DAI X, et al. Token merging: Your ViT but faster[C]//Proceedings of International Conference on Learning Representations (ICLR). Kigali, Rwanda: [s.n.], 2023.

[77] YUAN Z, XUE C, CHEN Y, et al. PTQ4VIT: Post-training quantization for vision transformers with twin uniform quantization[C]//Proceedings of European Conference on Computer Vision (ECCV). Tel Aviv, Israel: Springer, 2022: 191-207.

[78] LIN Y, ZHANG T, SUN P, et al. FQ-ViT: Post-training quantization for fully quantized vision transformer[C]//Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI). Vienna, Austria: Morgan Kaufmann, 2022: 1173-1179.

[79] LI Z, GU Q. I-VIT: Integer-only quantization for efficient vision transformer inference[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023: 17065-17075.

[80] LI X, LIU Y, LIAN L, et al. Q-Diffusion: Quantizing diffusion models[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023: 17535-17545.

[81] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: From error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[82] ZHANG R, ISOLA P, EFROS A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Utah, USA: IEEE, 2018: 586-595.

[83] DING K, MA K, WANG S, et al. Image quality assessment: Unifying structure and texture similarity[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(5): 2567-2581.

[84] SU S, YAN Q, ZHU Y, et al. Blindly assess image quality in the wild guided by a self-adaptive hyper network[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2020: 3667-3676.

[85] ZHANG W, ZHAI G, WEI Y, et al. Blind image quality assessment via vision-language correspondence: A multitask learning perspective[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023: 14071-14081.

[86] TALEBI H, MILANFAR P. NIMA: Neural image assessment[J]. IEEE Transactions on Image Processing, 2018, 27(8): 3998-4011.

[87] KIRSTAIN Y, POLYAK A, SINGER U, et al. Pick-a-Pic: An open dataset of user preferences for text-to-image generation [C]//Proceedings of International Conference on Neural Information Processing Systems (NIPS). New Orleans, USA: MIT

Press, 2023: 36652-36663.

[88] XU J, LIU X, WU Y, et al. ImageReward: Learning and evaluating human preferences for text-to-image generation[C]//Proceedings of International Conference on Neural Information Processing Systems (NIPS). New Orleans, USA: MIT Press, 2023: 15903-15935.

[89] ZHANG S, WANG B, WU J, et al. Learning multi-dimensional human preference for text-to-image generation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington, USA: IEEE, 2024: 8018-8027.

[90] MA Y, WU X, SUN K, et al. HPSv3: Towards wide-spectrum human preference score[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2025: 15086-15095.

[91] KOSTI R, ALVAREZ J M, RECASENS A, et al. Emotion recognition in context[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Hawaii, USA: IEEE, 2017: 1667-1675.

[92] POSNER T, FEI-FEI L. AI will change the world, so it's time to change AI[J]. Nature, 2020, 588(7837): S118.

[93] LIANG W, TADESSE G A, HO D, et al. Advances, challenges and opportunities in creating data for trustworthy AI[J]. Nature Machine Intelligence, 2022, 4(8): 669-677.

[94] HUANG P, DING W, STOLER B, et al. CaDRE: Controllable and diverse generation of safety-critical driving scenarios using real-world trajectories[C]//Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Georgia, USA: IEEE, 2025: 5474-5481.

[95] PALEJA R, MUNJE M, CHANG K, et al. Designs for enabling collaboration in human-machine teaming via interactive and explainable systems[C]//Proceedings of International Conference on Neural Information Processing Systems (NIPS). Vancouver, Canada: MIT Press, 2024: 64942-64969.

[96] CHAKRABARTY P K. Causal inference in agentic AI: Bridging explainability and dynamic decision making[J]. International Journal of Science and Research, 2025, 14(4): 2112-2117.

[97] ORZIKULOVA A, XIAO H, LI Z, et al. Time2Stop: Adaptive and explainable human-AI loop for smartphone overuse intervention[C]//Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI). Hawaii, USA: ACM, 2024: 1-20.

[98] KAZEMITABAAR M, WILLIAMS J, DROSOS I, et al. Improving steering and verification in AI-assisted data analysis with interactive task decomposition[C]//Proceedings of the Annual ACM Symposium on User Interface Software and Technology (UIST). Pittsburgh, USA: ACM, 2024: 1-19.

[99] BELLOS F, LI Y, SHU C, et al. Towards effective human-in-the-loop assistive AI agents[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2025: 2513-2522.

[100] GRANGE C, DEMAZURE T, RINGEVAL M, et al. The human-genAI value loop in human-centered innovation: Beyond the magical narrative[J]. Information Systems Journal, 2026, 36(1): 29-51.

[101] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 618-626.

[102] KIM B, WATTENBERG M, GILMER J, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)[C]//Proceedings of the International Conference on Machine Learning (ICML). Stockholm, Sweden: ACM, 2018: 2668-2677.

[103] CHEN C, LI O, TAO C, et al. Deep learning for interpretable image recognition[C]//Proceedings of International Conference on Neural Information Processing Systems (NIPS). Vancouver, Canada: MIT Press, 2019: 8930-8941.

[104] ZHANG Y, TIVNO P, LEONARDIS A, et al. A survey on neural network interpretability[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2021, 5(5): 726-742.

[105] SCHÖLKOPF B. Causality for machine learning[M]. New York, USA: ACM, 2022: 765-804.

[106] WU T, YANG G, LI Z, et al. GPT-4V (Vision) is a human-aligned evaluator for text-to-3D generation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington, USA: IEEE, 2024: 22227-22238.

[107] LI W, ZHANG R, SHAO R, et al. CogVLA: Cognition-aligned vision-language-action model via instruction-driven routing & sparsification[C]//Proceedings of International Conference on Neural Information Processing Systems (NIPS). Sydney, Australia: MIT Press, 2025.

[108] JI Y, TAN H, SHI J, et al. RoboBrain: A unified brain model for robotic manipulation from abstract to concrete[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Tennessee, USA: IEEE, 2025: 1724-1734.

[109] LI Z, REN L, YANG J, et al. VIP: Vision-instructed pre-training for robotic manipulation[C]//Proceedings of the International Conference on Machine Learning (ICML). Vancouver, Canada: ACM, 2025: 35769-35778.

[110] GAWLIKOWSKI J, TASSI C R N, ALI M, et al. A survey of uncertainty in deep neural networks[J]. Artificial Intelligence Review, 2023, 56(S1): 1513-1589.

[111] LI J, MIAO Z, QIU Q, et al. Training Bayesian neural networks with sparse subspace variational inference[C]//Proceedings of the International Conference on Learning Representations (ICLR). Vienna, Austria: [s.n.], 2024.

[112] ZEEVI T, SHWARTZ-ZIV R, LECUN Y, et al. Rate-In: Information-driven adaptive dropout rates for improved inference-time uncertainty estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Tennessee, USA: IEEE, 2025: 20757-20766.

[113] ABE T, BUCHANAN E K, PLEISS G, et al. Pathologies of predictive diversity in deep ensembles[C]//Proceedings of the International Conference on Learning Representations (ICLR). Vienna, Austria: [s.n.], 2024.

[114] RAMÉ A, AHUJA K, ZHANG J, et al. Model ratatouille: Recycling diverse models for out-of-distribution generalization[C]//Proceedings of the International Conference on Machine Learning (ICML). Honolulu, USA: ACM, 2023: 28656-28679.

[115] ZHU F, CHENG Z, ZHANG X Y, et al. OpenMix: Exploring outlier samples for misclassification detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023: 12074-12083.

[116] ZHANG H, ZHUANG Z, ZHAO H, et al. ReinboT: Amplifying robot visual-language manipulation with reinforcement learning[C]//Proceedings of the International Conference on Machine Learning (ICML). Vancouver, Canada: ACM, 2025.

[117] FANG H, GROTZ M, PUMACAY W, et al. SAM2Act: Integrating visual memory for robust robot manipulation under occlusion[C]//Proceedings of the International Conference on Machine Learning(ICML). Vancouver, Canada: ACM, 2025.

[118] HUANG C P, WU Y H, CHEN M H, et al. ThinkAct: Vision-language-action reasoning via reinforced visual latent planning [C]//Proceedings of International Conference on Neural Information Processing Systems (NIPS). Sydney, Australia: MIT Press, 2025.

[119] HSU C C, WEN B, XU J, et al. Spot: SE(3) pose trajectory diffusion for object-centric manipulation[C]//Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Yokohama, Japan: IEEE, 2025: 4853-4860.

[120] ZHENG J, LI J, LIU D, et al. Universal actions for enhanced embodied foundation models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Tennessee, USA: IEEE, 2025: 22508-22519.

[121] WAN W, ZHU Y, SHAH R, et al. LOTUS: Continual imitation learning for robot manipulation through unsupervised skill discovery[C]//Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Yokohama, Japan: IEEE, 2024: 537-544.

[122] SINGH A. Human-computer interaction: A review of usability, design, and accessibility trends[J]. Global Research Report, 2025(2): 362-387.

[123] HO M R, SMYTH T N, KAM M, et al. Human-computer interaction for development: The past, present, and future[J]. Information Technologies & International Development, 2009, 5(4): 1-18.

[124] SHACKEL B. Ergonomics in the design of a large digital computer console[J]. Ergonomics, 1962, 5(1): 229-241.

[125] LICKLIDER J C R, CLARK W E. On-line man-computer communication[C]//Proceedings of the Spring Joint Computer Conference(AFIPS). California, USA: ACM, 1962: 113-128.

[126] PREECE J, ROGERS Y, SHARP H, et al. Human-computer interaction[M]. Edinburgh Gate Harlow, UK: Addison-Wesley Longman Ltd., 1994.

[127] TOBY B H. EXPGU I. A graphical user interface for GSAS[J]. Applied Crystallography, 2001, 34(2): 210-213.

[128] WANG H, DING Q, LUO Y, et al. High-performance hydrogel sensors enabled multimodal and accurate human-machine interaction system for active rehabilitation[J]. Advanced Materials, 2024, 36(11): 2309868.

[129] PEI G, LI H, LU Y, et al. Affective computing: Recent advances, challenges, and future trends[J]. Intelligent Computing, 2024, 3: 0076.

[130] WANG J, YE W, HE J, et al. Integrating biological and machine intelligence: Attention mechanisms in brain-computer interfaces[J]. Information Fusion, 2026, 125: 103417.

[131] LAVALLE S M. Virtual reality[M]. Cambridge, UK: Cambridge University Press, 2023.

[132] DING X, HAN J, XU H, et al. Holistic autonomous driving understanding by bird's-eye-view injected multi-modal large models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Washington, USA: IEEE, 2024: 13668-13677.

[133] WEN L, FU D, LI X, et al. DiLu: A knowledge-driven approach to autonomous driving with large language models[C]//Proceedings of The Twelfth International Conference on Learning Representations(ICLR). Vienna, Austria: [s.n.], 2024.

[134] WANG S, YU Z, JIANG X, et al. Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Tennessee, USA: IEEE, 2025: 22442-22452.

[135] LI Y, TIAN M, LIN Z, et al. Fine-grained evaluation of large vision-language models in autonomous driving[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV). Paris, France: IEEE, 2025: 9431-9442.

[136] ZHANG R, ZHANG W, TAN X, et al. VLDrive: Vision-augmented lightweight MLLMs for efficient language-grounded autonomous driving[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV). Paris, France: IEEE, 2025: 5923-5933.

[137] ZHANG R, XIE J, ZHANG W, et al. Adadrive: Self-adaptive slow-fast system for language-grounded autonomous driving [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV). Paris, France: IEEE, 2025: 5112-5121.

[138] XIANG C G, YU Z. Human-machine hybrid augmented intelligence: Human-machine relationship, collaboration and mutual enhancement[C]//Proceedings of 2023 China Automation Congress (CAC). Chongqing, China: [s.n.], 2023: 7471-7478.

[139] AKATA Z, BALLIET D, DE RIJKE M, et al. A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence[J]. Computer, 2020, 53(8): 18-28.

[140] 国务院.关于印发新一代人工智能发展规划的通知:国发〔2017〕35号[EB/OL]. (2017-07-08)[2017-07-20]. https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm.
The State Council of the People's Republic of China. Notice of the state council on issuing the new generation artificial intelligence development plan (Guofa [2017] No. 35)[EB/OL]. (2017-07-08) [2017-07-20]. https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm.

[141] GOLDSTEIN E B. Cognitive psychology: Connecting mind, research, and everyday experience[M]. Stamford, USA: Cengage Learning Stamford, CT, 2015.

[142] SANFILIPPO F, ZAFAR M H, WILEY T, et al. From caged robots to high-fives in robotics: Exploring the paradigm shift from human-robot interaction to human-robot teaming in human-machine interfaces[J]. Journal of Manufacturing Systems, 2025, 78: 1-25.

[143] PARK J S, O'BRIEN J, CAI C J, et al. Generative agents: Interactive simulacra of human behavior[C]//Proceedings of the 36th Annual ACM Symposium on User Interface Software And Technology (UIST). California, USA: ACM, 2023: 1-22.

[144] METZGER S L, LITTLEJOHN K T, SILVA A B, et al. A high-performance neuroprosthesis for speech decoding and avatar control[J]. Nature, 2023, 620(7976): 1037-1046.

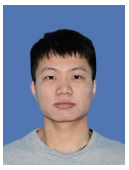
[145] CORVELO BENZ N, RODRIGUEZ M. Human-aligned calibration for ai-assisted decision making[C]//Proceedings of International Conference on Neural Information Processing Systems (NIPS). New Orleans, USA: MIT Press, 2023: 14609-14636.

[146] XINYU L, BINGKUN Y, PENGCHAO W, et al. An approach based on hybrid-augmented intelligence for the combination and optimization of human-machine teams[J]. Journal of Manufacturing Systems, 2025, 83: 306-321.

[147] ENDSLEY M R. Supporting human-AI teams: Transparency, explainability, and situation awareness[J]. Computers in Human Behavior, 2023, 140: 107574.
[148] PATRICK-KRUEGER K M, BURKHART I, CONTRERAS-VIDAL J L. The state of clinical trials of implantable brain-computer interfaces[J]. Nature Reviews Bioengineering, 2025, 3(1): 50-67.
[149] KENNEDY P R, BAKAY R A. Restoration of neural output from a paralyzed patient by a direct brain connection[J]. Neuroreport, 1998, 9(8): 1707-1711.
[150] HOCHBERG L R, SERRUYA M D, FRIEHS G M, et al. Neuronal ensemble control of prosthetic devices by a human with tetraplegia[J]. Nature, 2006, 442(7099): 164-171.
[151] HOCHBERG L R, BACHER D, JAROSIEWICZ B, et al. Reach and grasp by people with tetraplegia using a neurally controlled robotic arm[J]. Nature, 2012, 485(7398): 372-375.
[152] FLESHER S N, COLLINGER J L, FOLDES S T, et al. Intracortical microstimulation of human somatosensory cortex[J]. Science Translational Medicine, 2016, 8(361): 361ra141.
[153] BOUTON C E, SHAIKHOUNI A, ANNETTA N V, et al. Restoring cortical control of functional movement in a human with quadriplegia[J]. Nature, 2016, 533(7602): 247-250.
[154] WILLETT F R, AVANSINO D T, HOCHBERG L R, et al. High-performance brain-to-text communication via handwriting [J]. Nature, 2021, 593(7858): 249-254.
[155] WILLETT F R, KUNZ E M, FAN C, et al. A high-performance speech neuroprosthesis[J]. Nature, 2023, 620(7976): 1031-1036.
[156] LORACH H, GALVEZ A, SPAGNOLO V, et al. Walking naturally after spinal cord injury using a brain-spine interface[J]. Nature, 2023, 618(7963): 126-133.
[157] RUBIN D B, AJIBOYE A B, BAREFOOT L, et al. Interim safety profile from the feasibility study of the BrainGate neural interface system[J]. Neurology, 2023, 100(11): 1177-1192.

作者简介:

高新波(1972-),男,教授,博士生导师,研究方向:人工智能、机器学习、计算机视觉和模式识别等,E-mail:gaoxb@cqupt.edu.cn。



莫梦竟成(1997-),男,博士研究生,研究方向:具身智能,自动驾驶。

张灿(2001-),女,博士研究生,研究方向:图像超分辨。

袁钰(2000-),男,博士研究生,研究方向:目标检测。

张明珠(1998-),女,博士研究生,研究方向:低光照图像增强。

任路阳(2002-),男,硕士研究生,研究方向:行人重识别。

李爽(1995-),男,博士研究生,研究方向:行人重识别。

冷佳旭(1989-),**通信作者**,男,副教授,博士生导师,研究方向:人脸超分、行人重识别、视频异常检测、自动驾驶和具身智能,E-mail:lengjx@cqupt.edu.cn。

(编辑:陈珺)

Human-Centered Trustworthy Visual Intelligence

GAO Xinbo[1,2], MO Mengjingcheng[2], ZHANG Can[2], YUAN Yu[2], ZHANG Mingzhu[2], REN Luyang[2], LI Shuang[2], LENG Jiaxu[2*]

(1. School of Electronic Engineering, Xidian University, Xi'an 710071, China; 2. School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract: This survey reviews human-centered trustworthy visual intelligence by summarizing its application landscape, key techniques, and emerging trends. As computer vision advances from perception to highly autonomous decision making and physical execution, risks related to privacy, fairness, robustness, transparency, and safety become increasingly salient. When system outputs may affect human safety and rights, performance optimization alone can no longer satisfy the requirements for trustworthiness. From a computer vision perspective, the paper traces the concept and evolution of trustworthy visual intelligence, emphasizing the multiple roles of humans as data subjects, cognitive participants, and ultimate controllers. A unified framework is then presented along three complementary spaces, information, cognitive, and physical, and a progressive paradigm is formulated that focuses on humans, serves humans, and remains under human control. The survey synthesizes human-oriented visual data analysis methods under fairness and privacy constraints, robust and responsible model design strategies, and human-machine collaborative control mechanisms centered on transparency and safety, with discussions across representative scenarios such as image enhancement, video analysis, robotic manipulation, and 3D visual perception. Finally, open challenges and future directions are outlined, including robustness evaluation, cross-scenario generalization, collaborative governance, and sustainable deployment, providing a roadmap for trustworthy visual intelligence in real-world systems.

Highlights:

1. The paper proposes a human-centered unified framework for trustworthy visual intelligence, anchored by the triadic space of "information space, physical space, and cognitive space" and integrating the analytical chain from data to models to governance.

2. The paper develops an organizing scheme centered on "focusing on humans, serving humans, and remaining under human control". It aligns task and data object specification, model design with constraint injection, and explainable auditing with risk management, enabling a coherent narrative across application scenarios.

3. The paper outlines six principles for trustworthy deploying, including fairness, privacy, security, transparency, traceability, and robustness. It provides task-aware guidance that bridges model design choices with evaluation protocols to support systematic assessment and practical implementation.

Key words: trustworthy visual intelligence; human-centered; computer vision

---

Foundation items: National Natural Science Foundation of China (No.62472060); Chongqing Natural Science Foundation (Nos. CSTB2024NSCQ-QCXMX0060, CSTB2023NSCQ-LZX0061); Chongqing Key Research and Development Program of Science and Technology Innovation (No.CSTB2023TIAD-STX0016).

Received: 2026-01-28; **Revised:** 2026-02-28

***Corresponding author, E-mail:** lengjx@cqupt.edu.cn.