机器视觉编码技术研究及其进展

田港一[1),3)] 纪 雯[1),2)]

[1)](中国科学院计算技术研究所处理器芯片全国重点实验室 北京 100190)

[2)](龙眼国科智能信息技术有限公司 北京 100010)

[3)](中国科学院大学 北京 101408)

摘 要 机器视觉编码技术的不断进步,在工业制造、遥感监测、低空经济以及智慧安防等领域展现出广阔的应用前景。然而,现有的机器视觉编码技术和算法结构仍在不断演进,不同技术之间存在显著差异,导致研究人员对其理解还未达到一致,未能形成统一的标准和体系。因此,迫切需要对机器视觉编码技术的发展现状进行系统梳理,为未来研究和实际应用提供参考。本文从机器视觉的信源数据和编码原理入手,首先深入分析了机器视觉数据源的生成过程,包括信号采集和典型数据的编码特性等。接着,围绕"码率-任务质量-可计算性"的核心架构,阐述了机器视觉编码的基本原理,提出了通用的典型参考架构,并讨论了评价方法、可计算性分析、标准进展以及面临的技术瓶颈。针对机器视觉视频编码中的关键技术进行了全面论述,包括机器语义的视觉信息表示方法、特征生成及编码技术、面向任务的编码技术、可变码率优化技术以及近似重构技术。进一步,从编码器结构的角度出发,介绍了多种新型机器视觉编码器结构,如机器视觉的端到端编码结构、适用于可见光通用的混合编码结构、人机混合编码结构、集成编码结构、基于内容信息抽象表示的编码结构以及点云编码结构等。通过分析多种机器视觉数据类型的编码算法及其应用,详细评估了典型数据编码算法在不同应用场景中的优势与不足,并根据具体应用需求提出了系统性的解决方案。最后,对机器视觉编码技术的未来研究发展方向和应用前景进行了展望,旨在推动该技术在各领域的进一步发展和应用。

关键词 机器视觉数据源;机器视觉编码;编码结构;特征编码;深度学习

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2025.02631

A Survey on Machine Vision Coding Technologies

TIAN Gang-Yi[1),3)] JI Wen[1),2)]

[1)](State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

[2)](LonganPi Intelligent Information Technology Co., Ltd., Beijing 100010)

[3)](University of Chinese Academy of Sciences, Beijing 101408)

Abstract As machine vision coding improves by leaps and bounds, it demonstrates significant potential for applications in industrial manufacturing, remote sensing, the rapidly growing low-altitude economy, and smart surveillance. These fields require high-performance systems capable of processing large volumes of data in real-time. Unlike traditional video coding techniques that are optimized solely for human vision perception, machine vision coding is specifically designed for machine vision and various automated tasks. It eliminates redundant information that is not useful for automated processing. This technology primarily encodes the essential vision data necessary for machine tasks, thereby reducing data transmission costs. Furthermore, machine vision coding does not limit itself to the extraction of isolated features. Instead, it incorporates a broad

收稿日期:2024-06-18;在线发布日期:2025-04-10。本课题得到北京市自然科学基金(L221004)、国家重点研发计划(2023YFB4502805)、中国博士后科学基金(GZC20251091)以及江苏省工业人工智能重点实验室的资助。**田港一**,博士研究生,中国计算机学会(CCF)会员,主要研究领域为机器视觉编码、图像信号处理(ISP)、微光算法等。E-mail:tiangangyi22b@ict.ac.cn。**纪雯(通信作者)**,博士,研究员,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为视觉处理器、多媒体系统、工业人工智能,包括高性能视觉处理器、多媒体端边云计算系统、视觉编码与传输、工业智能芯片与系统等。E-mail:jiwen@ict.ac.cn。

range of data optimization strategies tailored to multiple machine vision applications. This approach minimizes reliance on conventional video content while preserving task-relevant information, ultimately enhancing the overall efficiency of machine operations. However, existing machine vision coding technologies and algorithms are still evolving. The technical roadmaps are still open. These differences have resulted in varying interpretations and approaches, hindering the development of a unified standard and framework for machine vision coding technologies. Therefore, this work provides a comprehensive, systematic survey of contemporary machine vision coding technologies, aiming to clarify the current development status and provide a reference for future research and practical applications. First, from the aspect of source data and fundamental coding principles in machine vision, we analyze the collection, generation, and coding characteristics of source data. Second, we propose a fundamental framework of "bitrate, task-quality, computability" and provide a reference architecture for machine vision coding, which guides the construction, optimization, and low-complexity design of the codecs. We also give evaluation methods, computability analysis, progress of standardization, and current bottlenecks. Additionally, we thoroughly analyze the key technologies of machine vision coding. It explores machine semantics-based vision representation to extract relevant data from complex scenes while minimizing irrelevant information. Feature generation and coding with efficient neural networks to support intelligent algorithms. Task-oriented coding is discussed for optimizing video efficiency by balancing accuracy and computational load. Variable bitrate optimization dynamically adjusts the bitrate based on network conditions. Approximate reconstruction techniques minimize data encoding and reduce transmission volume. Third, based on the perspective of codec architecture, we list typical codec structures, such as end-to-end coding for machine vision, hybrid coding for visible light data, human-machine hybrid coding, integrated encoding, content-based conceptual encoding, and point cloud coding. We give feasible systemic solutions through in-depth analysis of the strengths and weaknesses of those structures in different source data and practical applications. Finally, we highlight the ongoing research challenges and potential future directions, aiming to promote further development and application of this technology across various fields. Future research in coding can continue to explore several key areas, including the joint encoding of multi-sensor data, feature compression, semantic representation, human-machine hybrid coding, and the enhancement of scalability. Additionally, breakthroughs can be made in areas such as task-driven coding models and the development of vision processing units, which will improve coding efficiency and processing capability. Machine vision coding technology is currently a frontier research direction in the field of video coding.

Keywords source data of machine vision; machine vision coding; coding structure; feature coding; deep learning

1 引 言

面向机器视觉的编码技术(Machine Vision Coding)是近年来广受关注的研究热点。与传统针对人眼视觉优化的视频编码技术不同,机器视觉编码专为机器视觉和机器任务设计,去除机器视觉的冗余。主要通过对视觉数据中对机器任务关键的信息进行编码,以降低数据传输成本。机器视觉编码技术不仅限于特定信息的提取,而是涵盖了多种机器视觉任务所需的数据优化策略。该技术优势主要在于最小化对传统视频内容依赖的同时,保留任务相关信息,从而提升机器执行任务的效率[1]。

与传统编码技术相比,机器视觉编码的技术革新主要体现在以下四个方面:(1)编码目标的不同:传统视频编码面向人眼视觉,侧重于保留视频中的

图像细节,以实现最佳视频质量、分辨率和帧率的数据压缩[2]。而机器视觉编码面向机器任务,侧重于保留视频中的重要信息,提取机器任务所需的特征信息进行压缩编码,从而实现高效的机器任务执行;(2)冗余信息的处理:传统视频编码采用混合预测和变换设计方案,消除视频的时间冗余、空间冗余和统计冗余,生成可重建的、供人观看的数据。然而对于机器来说,这些编码后的视频流仍然含有大量"非有效信息冗余",需要进一步减少。机器视觉编码技术则可以显著减少这些冗余信息,提高压缩效率;(3)编码结构的灵活性:传统视频编码技术的核心模块固定,包括帧内/帧间预测、变换及量化、去块滤波器、熵编码等,导致编码端复杂度高。而机器视觉编码技术可以针对特定机器任务设计端到端神经网络结构、轻量化特征压缩模块,或为类似任务设计共性化特征压缩模块,以及人机混合应用设计兼容化模块,使编码端复杂度更低。此外,由于特征编码技术的引入,编码的特征信息无法恢复成人眼可见的视频,这对存储或传输过程中的数据安全具有较好的保密性;(4)核心问题的定义:传统视频编码的核心问题是码率－失真优化 RDO(Rate-Distortion Optimization),即在一定的码率约束下,最小化编码后的失真。机器视觉编码的核心问题可分为码率 R(Rate)、任务质量 Q(Quality)、可计算性 C(Computability)三个模块,即 R-Q-C 问题。其背后的基本原理涉及电磁波信号的采集、信号传感器、数据统计特性、编码框架、评价方法及应用瓶颈等多个方面。这些原理融合了电磁波信号、信源编码、机器视觉和通信等学科的基础理论,极具挑战性。

由机器主导的人工智能场景高速发展,面向机器视觉的编码需求急剧上升。在工业制造场景中,机器每秒采集与处理高达十亿甚至百亿级的像素[3-4]。由于存储、计算和电力资源有限,传统编码技术难以在大数据采集、快速反应和计算中应用过多保留了视觉细节。在遥感监测及低空经济领域[5],由于需要整合来自不同传感器的多波段数据以完成视觉任务,对机器视觉编码的抗差错性和可重复性提出了更高的要求。相比传统视频编码技术,机器视觉编码技术能够在语义质量方面提升了编码效率,提高了有效信息的获取精度和速度,大幅节省了人力资源成本。因此,机器视觉编码技术在工业制造、遥感监测、低空经济等需要机器视觉的工业应用领域,以及智慧城市、智能交通等基于多视角的安防领域,均有广阔的应用前景[6-7]。

机器视觉编码技术具有重要的学术价值和广阔的应用前景,国内外一些重要机构开展了专项研究。机器视觉编码技术源自 MPEG(Moving Picture Experts Group)专家组在 2015 年建立的视觉搜索紧凑描述子 CDVS(Compact Descriptors for Visual Search)标准[8],以及 2019 年建立的视频分析紧凑描述子 CDVA(Compact Descriptors for Video Analysis)标准[9]。这些标准采用紧凑特征描述视觉的比特流信息,实现了高效的图像和视频检索与分析。随后,MPEG-AI VCM(Video Coding for Machines)小组提出了一种针对人机混合视觉的联合视频编码方法,将视频编码分成两条码流,其中一条码流供机器视觉使用,两条码流共同解码给人眼观看。北京大学提出了规范的融合机器智能多任务的编码原型结构[10],考虑了包含视频、特征和模型在内的多个数据流协同的编码系统。中国科学院计算技术研究所团队提出多项关键技术:计算节点中深度学习模型通过特征提取和分辨率保持,优化传输函数减少视频数据[11-12];轻量级多注意力递归残差卷积神经网络编码方法利用多层次注意力机制提升编码性能与解码质量[13];高压缩比编码方法减少冗余信息引导机器深入理解低信息熵语义[14];屏幕内容感知视频编码方法将非局部模型与帧间滤波结合,提升编码效率和帧质量[15];软件可定义型视频编解码技术实现了视频编解码的灵活性和高效性[16]。针对机器视觉的特征编码方法、基于神经网络的端到端编码系统等研究也陆续展开[17-18],对机器视觉编码技术的发展和推广起到了重要推动作用。

目前,机器视觉编码技术的发展正在不断推进中。理论上,在同等视频任务完成质量下,仅面向机器的视觉编码性能可以远超传统视频编码。此外,编码的视觉数据源应覆盖机器能够使用的全波段数据,以充分利用各类传感器获取的重要数据。然而,相对于成熟的传统视频编码方法,如 H.26x 系列、AVS(Audio and Video Coding Standard in China)系列等[19],机器视觉编码在泛化性和稳定性上仍存在明显差距,编码的数据类型集中在可见光这一狭窄波段内,距离大规模应用还面临一些技术挑战,有待进一步深入研究。

由于机器视觉应用的首要痛点是执行高质量视觉任务所需的数据量巨大,因此 R-Q-C 问题指机器视觉编码需求是在保证视觉任务精度不小于阈值 Q、各项机器计算复杂度之和不大于阈值 C 的约束

条件下，使用最低编码码率 R 的编码方法。然而，机器视觉编码核心问题背后的基本原理尚未完全明确，缺乏可供参考的系统性研究论述。为了让更多学者了解机器视觉编码技术的研究现状，推动机器视觉编码技术领域的研究，本文的主要贡献如下：

(1)全流程系统性分析机器视觉编码技术：首次从输入到机器视觉数据源开始，对整个编码流程进行了系统性分析。通过深入解析数据采集、原理、编码、应用等环节，总结了不同数据特性的编码策略。与现有研究相比，本文提供了更为全面的视角，使得研究者能够更有效地了解从数据源头到最终应用的整个编码过程。

(2)提出典型的编码框架：在本文第 3.1 节中，提出了一种专为机器视觉设计的典型编码框架。在梳理既有研究成果的基础上，结合机器视觉的独特需求和技术限制，对现有编码理论进行了针对性优化。该框架为研究人员提供参考，有助于推动该领域的进一步发展。

(3)探讨机器视觉中非可见光区域的编码技术：在分析机器视觉领域主流的可见光编码技术的基础上，本文进一步拓展至非可见光区域，涵盖了点云、合成孔径雷达(SAR)、高光谱成像(HSI)等特殊数据的编码方法。通过梳理和分析这些技术的特点，本文总结了它们在特定应用中的优势与不足，提供了对这些领域编码方法的理解。

本文在第 2 节深入分析了机器视觉数据源，探讨了不同类型的机器数据采集过程和特点；第 3 节详细阐述了机器视觉编码的基本原理，并提出了通用的典型框架；第 4 节介绍了机器视觉编码中的关键技术，包括视觉表示、特征编码、任务编码技术等；第 5 节讨论了机器视觉编码结构，包括混合编码、端到端编码、人机混合编码等；第 6 节系统性地介绍了多种机器视觉数据类型的编码算法及其应用，包括可见光、点云、SAR 和高光谱数据；上述章节系统性分析了机器视觉编码的研究进展，在第 7 节对未来的研究技术点、发展方向和应用前景进行展望。

2 机器视觉数据源分析

机器视觉数据源分析是机器视觉技术中不可或缺的一环，信源数据的种类、模态分布特性和采集质量直接决定了机器视觉编码的覆盖范围和性能表现。相较于传统视频编码局限于人眼可见光范围，机器视觉传感器具备捕获更广泛电磁波信号的能力。数据源采集传感器作为数据获取的核心组件，目前捕获像素精度已高达上亿，确保了数据的价值与应用前景。对典型数据的深入剖析，有助于更深刻地理解数据源特性，进而优化处理算法，提升机器视觉系统的整体性能与精度。

2.1 电磁波信号

电磁波是由 James Clerk Maxwell 在 1865 年提出的重要物质，用于解决视觉感知中光的本质问题，并由 Heinrich Hertz 在 1888 年证实了电磁波的存在[20]，为机器捕获电磁波信号并将其解释为视觉数据奠定了理论基础。电磁波信号由空间中的交变电磁场产生，以波动的形式传播，其覆盖的频率范围非常广泛。机器能够捕获从低频到高频的各种波信号，处理人类看不见或看不过来的多维世界。受到鲨鱼、狗等不同生物感光细胞的启发，机器通过获取包括 γ 射线、X 射线、紫外线、可见光、红外线、无线电波等精细波谱来建立感知视觉系统，从而精确获取物质独特的特性表示。将电磁波按照它们的波长、频率或波数的顺序进行排列，形成如图 1 所示的信号源频率范围。

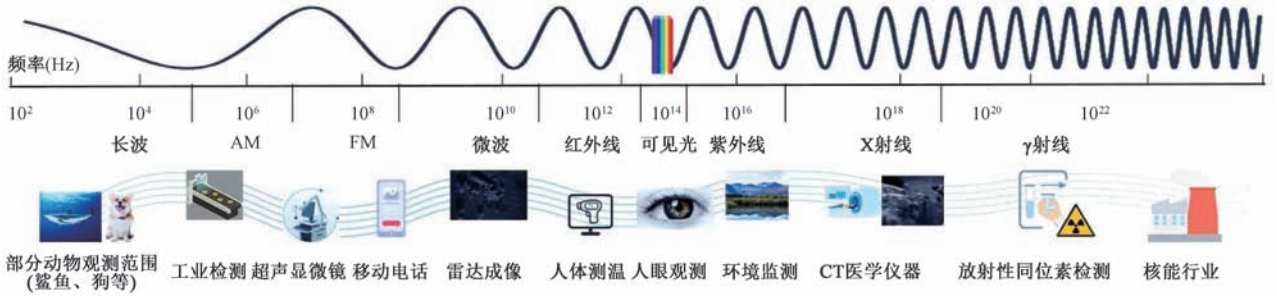


图 1 机器视觉的信号源频率范围

在电磁波段中，人眼视觉仅能感受可见光范围，波长约在 400 至 700 纳米的狭小区域。当前机器视觉系统中有许多针对可见光信号的人眼视觉图像和视频处理算法，但当这些算法用于处理可见光波长

以外的电磁波段频谱数据时,尤其是高频谱包含的庞大数据信息,使其难以进行快速的数字化存储和分析。在电源供电的移动终端设备应用场景下,带来了能耗成本问题。在实际应用中,机器视觉编码的信号范围包括用于超声显微镜、SAR 成像等低频段区域,红外人体测温、人眼观测、紫外线探测等中频段区域,以及 CT 医学检测、工业检测、地质石油探测等高频段区域。机器视觉编码技术有望从采集源头开始去除机器任务中冗余的频谱信号携带的信息,从而减轻处理器的计算负担,这对提升机器的视觉感知与处理效率具有重要作用。

2.2　数据源采集传感器

用于机器视觉的采集传感器将电磁波信号或光源信号转换成图像等视觉信号,发送到专用的视觉处理系统,并根据像素分布、亮度和颜色等信息将其转换成数字信号。根据采集信号的不同,机器视觉传感器可分为无源和有源两种类型。

在可见光的波长范围内,采集传感器使用无源器件,主要依靠 CCD(Charge Coupled Device)或 CMOS(Complementary Metal Oxide Semiconductor)两种芯片接收光强信息,但不能识别波长信号。因此,无源采集传感器的工作环境依赖于外部光源,在昏暗环境下的工作能力大幅下降。

在可见光的波长范围之外,采集传感器使用有源器件,这些器件能够自行发送和接收电磁波信号进行视觉感知,并且不受外部光源影响。相反,有源传感器需要消除同频干扰以达到更好的感知效果。这种差异使得有源采集传感器在某些应用场景下更为可靠。例如,在恶劣天气、灾害应急、航空遥感等场景下,可见光信号的传输受到极大限制,而有源采集传感器可以在这些条件下仍然有效地进行视觉感知。此外,有源采集传感器还可以在更长的距离范围内进行探测,并且能够感知到更小的目标。因此,在需要高精度和高可靠性的视觉感知任务中,有源采集传感器具有明显的优势。

采集传感器经过多年的发展,其性能不断提升。从 20 世纪 50 年代出现的光学倍增管,到 20 世纪 80 年代的 CMOS 无源采集传感器,再到 2000 年后的激光雷达和光场相机等有源采集传感器,图像传感器的发展历程中,每一代传感器都在不断提升像素采集的精度和帧率捕捉的准确度[21]。现今,民用级传感器的像素采集数已轻松突破亿级,而在工业级及特定应用场景下,传感器的像素数更是高达十亿以上。随着采集传感器性能的不断提升,带来了更加精准和高清的视觉体验,同时也为编码技术带来了超高码率的挑战。

从采集传感器的数目来看,多目采集传感器可以提供更丰富的视觉信息,特别是人眼视觉难以直接获取的深度、视差、全景等信息,从而更准确地感知和理解环境。从波段来看,不同波段信号,尤其是面向高频电磁波段获取一些人眼无法观察到的信息,如温度、热量、化学成分等,也能大幅提升机器在智能识别和分析方面的优势。传感器采集的数据越多,机器视觉的能力越强,但同时也增加了机器计算的负担。因此,对于机器视觉而言,合适的编码方法非常重要。传统的编码方法难以迅速处理多目、多波段数据,这限制了机器视觉在更长波段(如太赫兹)上的应用。为了突破传统编码方法的限制,机器视觉编码成为一种有效技术。通过机器视觉编码,机器可以更快速、更准确地处理高信息含量的数据,从而实现更广泛的应用。

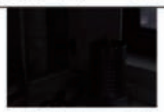
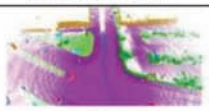
2.3　典型数据分析

传感器采集的电磁波信号的数据类型高达到数十种,不同类型的数据具有不同的信号特点和统计特点[22-28],如图 2 所示。本文选取几种典型的数据进行分析:

(1) 可见光数据:光学传感器将可见光信号转换为电信号,采集的可见光数据能够反映物体的动态变化和表面纹理。其波长范围大约在 400 至 700 纳米之间,覆盖了人眼可见的整个光谱段。空间分辨率通常较高,像素级细节适合图像识别与计算机视觉任务。动态范围为标准的 8～12bit,能够较好地处理亮度差异。穿透能力较弱,仅能在光线较强的环境中有效工作。实时处理性较高,支持每秒 60 帧的高速成像。光谱维度由 3 个可见光通道组成。可见光数据通常具有较高的时间相关和空间相关的编码统计特性,适用于图像识别和计算机视觉任务。在不同光照条件下,其编码策略可以通过视频快照压缩成像系统实现,以确保在变化的光照条件下保持图像稳定性[29]。在暗光环境中,编码方法需要考虑压缩效应对低光图像的影响,并恢复丢失的信息[30]。

(2) 红外数据:通过红外传感器或红外相机进行采集,这些传感器能够接收并转换物体发出的红外辐射为电信号。红外数据能够提供物质表面和内部分子结构信息。其波长范围通常大于 700 纳米,能够探测到物体发出的热辐射。空间分辨率一般较中等,通常为 640×512 分辨率,受热扩散效应限制。动态范围为 14 bit,能够捕捉较大的亮度差异,适合温度变化检测。穿透能力较弱,红外数据仅能穿透

薄雾。实时处理性较中等，适合中频采集（10～30 Hz）。光谱维度为单波段热辐射数据。由于不受光照条件的影响，红外数据适用于在夜间或光线较暗环境下的视觉任务。红外数据的编码统计特性包括高动态范围，即包含大量亮度差异较大的像素值，并且物体通常具有平滑的表面和渐变的温度变化[31]。

波长 λ (m)	数据类型	示例	数据特点	
			优势性	局限性
4×10^{-7} 至 7×10^{-7}	RGB	 户外场景下的RGB数据[22]	● 丰富的颜色和细节信息 ● 易于获取，具有实用性和可用性 ● 在多领域的应用广泛	● 光照变化导致亮度对比波动大 ● 阴影干扰影响机器视觉任务 ● 难以处理透明和反射材质
		 暗光场景下的RGB数据[23]		● 光线不足噪声增加，细节模糊 ● 亮度受限无法兼顾明暗细节 ● 颜色失真色准低，信息失效
7.6×10^{-7} 至 1×10^{-3}	红外线	 夜间的热红外数据[24]	● 检测物体的温度分布 ● 在黑暗环境下提供清晰的图像	● 受环境温度和湿度等因素干扰 ● 探测距离和探测角度有限
4×10^{-7} 至 2.5×10^{-6}	高光谱	 航空高光谱数据[25]	● 大面积连续覆盖成像 ● 检测潜在环境问题 ● 精确物质分类识别	● 需要复杂的信号处理和数据分析 ● 受到天气条件影响
1×10^{-3} 至 1×10^{-2}	点云	 自动驾驶点云数据[26]	● 提供大规模的三维空间信息 ● 不受视角影响	● 具有稀疏性和不规则性 ● 需要大量的计算资源处理数据
$\leqslant2\times10^{-2}$	超声波	 医疗腹部超声扫描数据[27]	● 反应速度灵敏，可实时成像 ● 超声波成像的距离远	● 传播过程中会受到介质的影响 ● 传输速度易受天气情况的影响
2.5×10^{-2} 至 1	合成孔径雷达SAR	 SAR低空数据[28]	● 提供高分辨率的地形测绘数据 ● 能够穿透云层与植被成像	● 成像的数据具有几何畸变 ● 数据处理复杂耗时

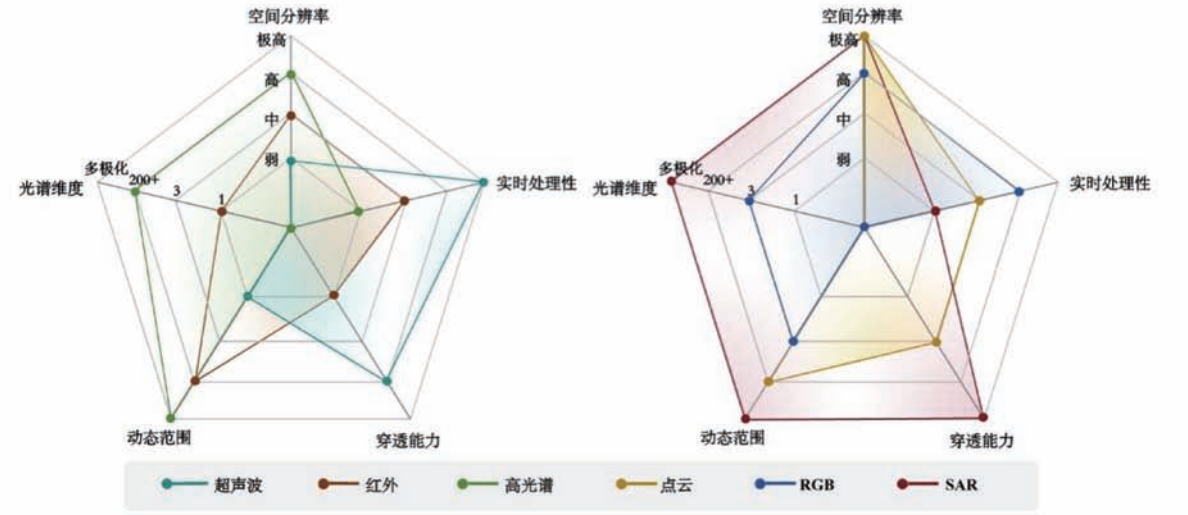


图 2 （数据特点分析表、六种典型数据的五维特性对比图）（空间分辨率、动态范围、穿透能力、实时处理性采用弱/中/高/极高评级。光谱维度采用通道特征描述，包括 1(单通道)、3 通道、200＋通道、多极化。）

(3) 高光谱数据:高光谱数据具有丰富的光谱信息,能够提供物质的细致分类和成分分析。这些数据在多个波段内进行采集,波长范围通常在 400 至 2500 纳米。空间分辨率为 1-5MP 像素,能够提供较高的细节信息。动态范围为 16 bit,支持高精度的亮度捕捉。穿透能力较弱,受限于只能穿透薄雾。实时处理性较低,通常为 0.1 帧每秒(fps),适用于低频扫描。光谱维度非常丰富,通常包含 200+个通道。高光谱数据的编码特点包括高维数据稀疏性和相关性较强,适用于复杂场景的目标识别和变化检测。

(4) 点云数据:激光雷达、毫米波雷达等设备发射激光脉冲并测量其反射回来的时间,从而计算三维空间中的距离信息,组成点云数据,反映物体的三维空间信息和形状特征。空间分辨率通常较高,能够提供细致的空间分布。动态范围较广,深度信息具有 14bit 量化,能够精准捕捉不同深度的细节。穿透能力较中等,点云数据能在有限条件下穿透雨雾(30%的穿透率)。实时处理性中等,适合中频采集,光谱维度为单波段。点云数据的波长范围在 905 至 1550 纳米,编码中具有较高的稀疏性和非均匀性,广泛应用于三维重建、自动驾驶、地形测绘等领域。

(5) 合成孔径雷达 SAR 数据:SAR 数据通过微波信号获取目标的高分辨率图像,具有穿透性强和全天候成像的特点。空间分辨率可以达到亚米级,通常为 0.1m。动态范围极高,能够达到 60dB 的动态范围。穿透能力极强,能够穿透云层和植被,尤其在 L 波段,穿透深度超过 10m。实时处理性较弱,SAR 数据的采集通常为低频扫描(每分钟采集一次)。光谱维度具有全极化模式,含 4 个偏振通道,能够获取多维信息。SAR 图像常受到干扰和噪声影响,因此在编码时需考虑对抗干扰的策略,以确保图像的准确性和可靠性。

(6) 超声波数据:超声波数据通过发射高频声波并接收其反射波来获取物体的结构信息。超声波传感器在医学、工业检测等领域广泛应用。空间分辨率受波长限制,通常为 1mm 级别,适用于精细的内部结构检测。动态范围较低,通常为 8bit 量化,适合较为简单的亮度和深度变化检测。穿透能力较强,能够穿透生物组织,常用于医学成像(如腹部超声穿透 20cm)。实时处理性非常高,支持每秒 30 帧的实时流处理,适合快速变化环境中的动态监测。光谱维度为单通道,通过接收声波反射信号来获取物体的状态信息。超声波数据通常具有较高的实时处理能力和较好的穿透能力,但在空间分辨率和动态范围方面存在一定限制,适用于细节不那么复杂的环境中的快速检测和诊断。

针对这些数据具有各自独特的信号特点和统计规律,研究机器视觉编码技术至关重要。虽然可见光领域的机器视觉编码技术已有所进展,但红外和超声数据等在编码技术研究方面仍显不足。针对当前研究热点的可见光、高光谱、点云和 SAR 数据,本文将在第 6 节深入探讨机器视觉编码算法及其应用。

3　机器视觉编码的基本原理

对于机器视觉编码的核心 R-Q-C 问题,本节将按照机器视觉编码的过程,逐层剖析核心问题背后的原理。首先,从机器视觉编码的框架来阐述码率 R。接下来,在机器视觉编码的评价方法小节中说明任务执行质量 Q。可计算性 C 主要关注编码数据量问题。此外,还探讨了机器视觉编码领域的最新发展和标准化趋势。最后,在技术瓶颈模块中讨论当前机器视觉编码面临的技术难题和限制因素。

3.1　机器视觉编码框架

自 1962 年第一颗电视直播卫星发射成功以来,第一代视频编码技术开始出现,主要解决人眼观看的清晰度和压缩率问题。随着 2000 年网络技术和城市建设的发展,第二代视频编码技术重点解决人眼视觉体验和弹性传输问题。到 2013 年,物端智能机器设备的数量快速增长,随之出现了第三代视频编码技术,需要重点解决的主体从人眼视觉逐步转向机器视觉[32-33]。

由于人眼视觉系统和机器视觉系统存在本质差异,仅面向人眼视觉的传统编码方法已经不能满足机器的视觉任务需求。人眼仅对可见光波段的颜色、亮度和对比度敏感,但对细节和纹理感知较弱。传统视频编码技术利用这些特点,消除时间、空间和统计冗余,保留图像细节。尽管视频质量、分辨率和帧率有所提高,但对机器视觉任务至关重要的信息密度却可能并未相应增加。这种情况下,视频流中仍然包含大量对机器而言的冗余信息。

研究表明[34],当采用高压缩率编码(如 QP>45)时,机器视觉模型的识别准确率可能下降达 30%,但例如保留 5%的关键纹理特征时也能正确识别目标。所以,可以通过选择性保留机器视觉敏

感特征来节省比特，找到最佳编码方式。机器视觉系统需要结构化解析图像/视频内容，以提取任务所需的特征信息[35]。因此，机器视觉编码技术应减少机器不可感知的语义冗余，以降低码率并提升任务执行质量。

本文提出的典型机器视觉编码框架如图 3 所示。相比于对全部原始视觉数据的有损或无损编码方案，机器视觉的编码过程如下：

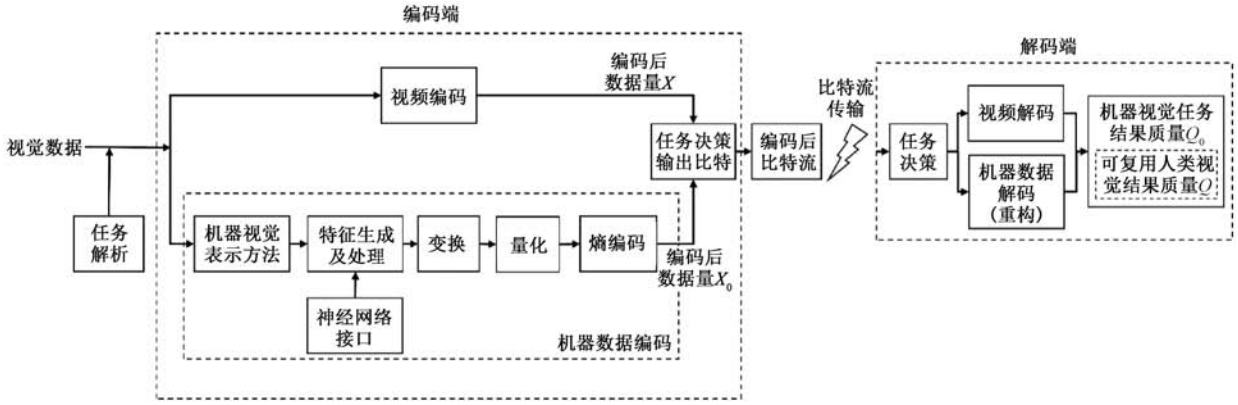


图 3 机器视觉编码典型框架

在编码端，原始视觉数据经过任务解析模块，判断是仅使用传统视频编码器得到码流 X，还是输入到机器视觉语义表示模块，以转换成更适用于任务的表示形式。经过表示方法模块处理后的数据通过机器学习方法（如神经网络模型）生成特征或感兴趣区域 RoI（Region of Interest）等有效信息。最后，对生成的有效信息进行变换、量化和熵编码等操作，得到机器视觉编码后的码流 X_0。

在解码端，通过解析码流中的任务决策比特位，可以决定机器数据码流的解码方法，以及是否需要进行传统视频码流的解码。通过解码或重构后的数据，可用于机器视觉任务和人类视觉任务。机器视觉编码的目的是使编码后的数据量 $X_0 \ll X$，同时确保执行的视觉任务质量 $Q_0 \geqslant Q$。

3.2 评价方法

根据不同的机器视觉任务，其评价方法应具有多样性。传统编码通常以码率－失真优化 RDO 问题为基础，可表示为

$$J = R + \lambda \cdot D \tag{1}$$

其中，包含比特率 R、失真 D 和联合成本 J。λ 是拉格朗日乘子，用于平衡比特率和失真。

传统编码方法在评估时没有考虑到视觉数据的最终接收者是机器而非人眼。编码器的目标并非重建输入图像，而是提供良好任务性能的数据张量[36]。因此，应将任务损失纳入评价计算中。机器视觉编码的评价方法应综合考虑比特率、失真、任务性能以及压缩效率等多个方面，评价函数如下：

$$O_{\text{total}} = w_{\text{rate}} O_{\text{rate}} + w_{\text{mse}} O_{\text{mse}} + w_{\text{task}} O_{\text{task}} \tag{2}$$

其中，O_{rate} 代表编码器的压缩率或比特率相关的损失项。O_{mse} 表示均方误差 MSE（Mean Squared Error）损失，常用于衡量重建误差或失真程度。O_{task} 代表与特定机器视觉任务相关的损失项。w_{rate}、w_{mse} 和 w_{task} 分别是损失项的标量权重。

具体来说，评价指标主要包含：

（1）评价编码效率：每像素比特数 BPP（Bits Per Pixel），即衡量机器视觉编码或压缩过程中每个像素所需的平均比特数。计算公式如下：

$$BPP = \frac{\text{总比特数}}{\text{总像素数}} \tag{3}$$

在高光谱数据编码中，用每个波段每像素比特数 BPPPB（Bits Per Pixel Per Band）来衡量编码效率，反映每个波段中每个像素所需的平均比特数。

（2）评价编码质量：峰值信噪比 PSNR（Peak Signal-to-Noise Ratio）[37]，计算最大值信号和背景噪音的比例：

$$PSNR = 20 \cdot \log_{10}\left(\frac{MAX_I}{\sqrt{MSE}}\right) \tag{4}$$

其中，MAX_I 是可能的最大像素值，MSE 是均方误差，该公式计算原始和压缩后的视觉数据之间的均方误差。

结构相似性指数 SSIM（Structural Similarity Index）[38]，用于衡量两幅图像相似度的指标。计算公式如下：

$$\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{5}$$

其中，x 和 y 分别代表未经编码的无失真视觉数据和编码后的视觉数据，μ_x 和 μ_y 是它们的均值，σ_x^2 和 σ_y^2 是它们的方差，σ_{xy} 是它们的协方差，C_1 和 C_2 是为了避免分母为零而添加的小常数。

(3)评价机器视觉任务:用于衡量编码器在处理任务时的准确性和性能水平的评价指标各有差异[39]。在常见的目标检测任务中,评价指标包含平均精度均值 mAP(mean Average Precision)。通过计算精确率 P(Precision)和召回率 R(Recall),绘制 P-R 曲线并计算曲线下的面积来评估性能。精确率表示正确检测到的目标数量与所有检测到的目标数量之比,而召回率表示正确检测到的目标数量与所有真实目标数量之比。同时,特定 IoU 阈值(如 IoU=0.5 时的 AP50)下的平均精度也是重要的评估指标,它进一步细化了对检测精度的考量。除此之外,还可以采用 Top-k (如 Top-1 或 Top-5)准确率作为补充评价指标,它关注的是模型预测结果中置信度最高的前 k 个目标中,正确目标的比例,这对于需要快速响应或资源受限的应用场景尤为重要。在多目标跟踪任务中,评价模型准确性的关键指标则是 MOTA(Multiple Object Tracking Accuracy),公式为

$$MOTA = 1 - \frac{\sum_t (FN_t + FP_t + IDS_t)}{\sum_t GT_t} \quad (6)$$

其中，t 是时间步，FN_t 是在 t 的漏检数，FP_t 是在 t 的误检数，IDS_t 是在 t 的切换数，GT_t 是在 t 的真实目标数。

其他视觉任务,如面部特征点检测任务,使用平均归一化均方根误差 NRMSE(Normalized Root Mean Square Error)作为评估检测精度的关键指标,其低值反映了更高的定位准确性,是衡量面部特征点检测算法性能的重要标准。对于图像去斑任务,有效噪声水平 ENLs(Effective Number of Levels)作为评价指标,侧重于衡量算法在保留图像细节与去除噪声之间的平衡能力,高 ENLs 值表明算法在去除噪声的同时能更好地保留图像细节,是图像去斑算法优化的重要方向。

(4)其他评价指标:主要包括编码时长、算法参数量以及多任务泛化性等。编码时长是衡量编码器处理数据速度的关键指标,对于需要高效处理大量数据的应用场景至关重要。算法参数量则直接关联到算法的内存占用、训练成本以及模型复杂度,是评估模型资源消耗情况的重要依据。编码时长和参数量是反映计算复杂度的重要指标。此外,多任务泛化性指编码器在处理不同种类或不同领域的视觉任务时,能够保持或提升性能的能力。多任务泛化性强的编码器不仅能够适应更广泛的应用场景,还能通过共享特征表示和知识迁移,提高整体学习效率和效果。

3.3　可计算性分析

机器视觉任务需处理的数据量远超人眼视觉可处理范围[40]。人眼视觉可处理的数据量约为每秒 7.2 亿像素[41],而机器视觉系统常常需要处理更高的数据量。例如,中国科学院长春光学精密机械与物理研究所的五亿级像素云相机系统在国防、安防和民用等领域也有广泛应用。国外的美国 SLAC 国家加速器实验室相机甚至能够拍摄 32 亿像素的图像,用于太空探测。这些数据量庞大的图像和视频数据给机器视觉编码带来了巨大的挑战。

机器视觉编码需要满足更高的分辨率要求[42]。以 8 K 广播为例,其传输带宽为 100 Mbps。而按照视网膜分辨率的要求,即使采用 500 倍的压缩比,信道宽度也将达到约 5.2 Gbps,这远远超出了现有技术的处理能力。高分辨率图像和视频经过编码后还需要进行复杂的数据处理和算法计算。例如,图像和视频的压缩、分割、特征提取、运动估计等操作都需要大量的计算资源和算法支持。所以,机器视觉编码的计算复杂度非常高,不仅涉及编码过程本身,还包括对编码结果的解码和处理过程。

3.4　标准进展

本小节从国际、国内和非可见光编码标准三个方面介绍了近几年编码标准的发展进程。

(1)国际标准:近年来,机器视觉编码技术取得了显著进展。ISO/IEC MPEG 专家组早在 2015 年便提出了 CDVS 标准[8],该标准通过紧凑的视觉描述符实现了图像的高效检索与分析。随后,2019 年建立的 CDVA 标准[9]进一步拓展了视频分析的范围,提供对视频内容的紧凑描述,为机器视觉任务提供了重要支持。在此基础上,MPEG-AI VCM 小组自 2019 年起开始研究针对人机混合视觉的联合视频编码技术,并在 2024 年正式立项该标准。该标准基于神经网络的编码与解码流程,同时实现对人类视觉和机器分析任务的适配。通过前处理阶段的网络检测、感兴趣区域提取、空间/时间重采样等技术优化输入数据,再经过内置编码器进行层次化编码和基于神经网络的编码,生成可供机器和人眼分别处理的码流。解码端则恢复视频,并根据机器任务

的需求调整输出，确保编码和解码在不同应用场景中的适应性[43]。

在传统视频编码标准方面，2021 年推出的 H.266/VVC(Versatile Video Coding)标准[44]在全球范围内引领了视频技术的前沿，支持高达 16K 的超高清编码。作为该标准的进一步优化，Fraunhofer Heinrich-Hertz-Institute 开发了 ECM(Enhanced Compression Model)编解码器[45]，显著提升了压缩效率。同时，AOM(开放媒体联盟)正主导开发下一代 AV2 标准，以应对 8K 视频在主流流媒体平台上的应用需求，如 YouTube 和 Netflix。

(2)国内标准：在新型编码结构方面，2024 年 5 月发布的团体标准《面向机器智能的数据编码第 2 部分：图像》，是我国在基于神经网络的图像编解码技术方面取得重要进展。该标准的特点是普适多种可见光机器视觉任务，为数据密集型的智能应用提供高效编码支持。

在传统视频编码标准领域，2024 年发布的 AVS3 国家标准[46]推出了面向 8K 及 5G 产业应用的视频编码技术。与单纯追求超高分辨率编码的方案不同，AVS3 标准更侧重于兼顾人眼视觉和机器视觉的需求。与此同时，2024 年发布的《视频浅压缩编码第 1 部分：超高清视频分层编码》不仅支持超高清视频传输，还为实现复杂机器视觉任务奠定了坚实的技术基础。此外，最新 GY/T 412-2024《超高清视频图像质量客观评价方法》规定了适用于 4K/8K 超高清视频的全参考与无参考图像质量客观评价方法。

(3)非可见光编码标准：目前的主流视频编码标准主要集中于可见光数据的压缩处理，尽管这些标准在机器视觉中发挥了重要作用，但它们仍然无法满足不断增长的多样化数据需求，特别是在红外、紫外以及点云数据的编码领域。对于点云数据，MPEG 于 2020 年发布的 V-PCC(Video-based Point Cloud Compression)和 G-PCC(Geometry-based Point Cloud Compression)标准[47]，支持 3D 点云数据的高效压缩，为沉浸式体验，如 72K×36K 分辨率的全景视频编码奠定了基础。然而，在更高分辨率和更复杂的场景中，这些标准仍需进一步优化，以满足未来沉浸式视频和虚拟现实应用的需求。对于红外数据，虽然国家标准 GB/T 33773-2017 规范了音视频设备红外线遥控的编码规则，但目前针对机器视觉任务的红外数据编码标准尚处于初期阶段，提案数量有限。这表明在红外数据的编码标准化方面仍有许多工作需要完成。同样，紫外线、射线等其他非可见光数据类型在机器视觉中的应用也日益增长，然而相应的编码标准仍有待研究。

尽管在国际和国内编码标准方面取得了诸多进展，仍存在一些关键问题。首先，神经网络编解码效率亟待提升，以降低对硬件资源的高要求，尤其是在大规模应用中的兼容性和高效性问题。其次，现有标准需更好地适应多样化的数据类型和应用场景，特别是在终端设备性能受限的情况下，确保跨场景的普适性。此外，针对红外、紫外等非可见光数据的编码标准化仍处于早期阶段，未来应加快多模态数据的编码研究，完善标准框架，提升压缩效率，推动复杂机器视觉任务的广泛应用。

3.5 技术瓶颈

机器视觉编码在不同应用场景中面临着各种挑战和瓶颈。从码率 R、任务质量 Q、可计算性 C 三个维度进行分析：

(1)码率 R 层面：传统视频编码受分辨率、帧率和信号范围的影响，在可见光范围内生成的码率通常能够满足传输要求；而机器视觉编码则涉及更高分辨率的视频、动态点云数据和多频谱信号，生成的数据速率极高，难以实现高效传输与应用。例如，在自动驾驶和工业检测等场景中，采用 8K@30 fps 的未编码视频，其码率约为 23900 Mbps；在自动驾驶或机器人导航等动态场景中，车辆往往配备 2 至 4 颗激光雷达以实现全方位环境监测。单个激光雷达每帧生成约 10 万个点(每点 16 字节)，在 30 fps 下的码率约为 384 Mbps。当搭载 4 颗激光雷达时，整体数据量将显著增加，总码率将达到约 1536 Mbps；此外，在农业监测或医学成像中，多频谱成像系统若配置 5 个波段、每个波段采用 4K@60 fps，其单波段码率约为 11944 Mbps，总码率则达到约 59720 Mbps。

(2)质量评价 Q 层面：在传统视频编码中，评价标准主要关注编码效率和压缩比，追求清晰画面和高帧率。然而，机器视觉编码面临多样复杂的需求，对应着不同的评价标准。机器视觉任务的评价标准包括分类、目标检测、语义分割、目标跟踪、特征点检测和图像去斑等多种任务，难以统一化评价标准。

(3)可计算性 C 层面：传统视频编码的计算复杂度在可控范围内，适合消费级应用。机器视觉编码的计算复杂度较高，难以支持工业级应用。这意味着在实际应用中可能会面临计算资源限制和性能瓶颈。

4　机器视觉编码中的关键技术

机器视觉编码相较于传统的混合编码技术(如AVS、H.26x 等)[48],最主要的区别在于利用机器视觉语义表示和任务解析等独有的关键技术,提高编码效率和任务执行质量。通过对图 3 所示的整个编码流程进行分析,关键技术主要包括:基于机器语义的视觉信息表示方法、特征生成及编码技术、面向任务的编码技术、可变码率优化技术以及近似重构技术。本文还对机器视觉的主要编码算法按照其结构分支和主要创新的关键技术进行分类,可在表 1 中进行快速查找,并在表 2 中查看部分算法的分析。下面对这些关键技术分别详细介绍。

表 1　机器视觉编码主要算法索引

数据	结构分支	主要创新技术	涉及创新点	文　献
可见光	端到端编码结构	机器视觉语义表示、特征生成及编码技术	可微分语义分割引导的视觉编码框架	[11-12, 23, 29-30, 62-63, 67, 71,97-100]
		机器视觉语义表示、面向任务编码技术	时空语义蒸馏的紧凑表示方法	[1,61,70,82,94,101-102]
		特征生成及编码技术、可变码率优化技术	分层潜在表征生成技术,自适应量化熵编码	[18,36,86-87,103-105]
		近似重构技术	基于扩散模型的迭代式精细重构	[35,40,93,106-110]
	混合编码结构	机器视觉语义表示、特征生成及编码技术	基于深度特征分解的视觉表示方法,自适应上下文建模的编码技术	[52,56,111-112]
		特征生成及编码技术、可变码率优化技术	时空特征解耦技术,动态感知码率控制机制	[7,15,90-91,113-115]
		特征生成及编码技术、面向任务编码技术	多任务特征蒸馏框架,面向任务熵率约束编码	[79,88,91]
		可变码率优化技术	基于强化学习的动态码率分配策略	[34,116-118]
		近似重构技术	多尺度残差注意力重构,感知失真联合优化	[13, 19, 44-45, 81, 92, 116, 119-124]
	人机混合编码结构	机器视觉语义表示、特征生成及编码技术	视觉－符号双模态联合表示架构	[10,68,125]
		机器视觉语义表示、面向任务编码技术	语义抽象图谱引导的编码优化	[59,78,80,126-128]
		可变码率优化技术	人机协作的码率自适应调节机制	[33,88]
	集成编码结构	特征生成及编码技术、面向任务编码技术	多粒度特征交互编码框架,任务驱动特征选择	[14,43,57,64,129-131]
		特征生成及编码技术、近似重构技术	解耦式特征生成与渐进式重构	[2,32,60,118]
	内容表示编码结构	机器视觉语义表示、近似重构技术	层次化特征分解表示,语义引导的生成式重构	[50-51,53,132-134]
		机器视觉语义表示、特征生成及编码技术	提取图像中的高级视觉特征,转化为关键语义信息,实现紧凑编码	[58,72,135]
	其他	机器视觉语义表示	神经辐射场增强的视觉表示	[6, 17, 27, 49, 54-55, 65-66, 69,73-76,82,136-142]
		特征生成及编码技术	可变形卷积增强的特征提取网络	[16,83-85,115,143-147]
		面向任务编码技术	多任务联合优化的元学习编码框架	[3-5,77,148-149]
		近似重构技术	基于 StyleGAN 的感知驱动重构	[95,150-154]
点云	端到端编码结构	特征生成及编码技术、近似重构技术	稀疏体素上下文建模,密度自适应编码	[155-160]
	点云编码结构	机器视觉语义表示、特征生成及编码技术、近似重构技术	几何－颜色解耦表示,八叉树注意力编码	[161-167]
SAR	端到端编码结构	机器视觉语义表示、特征生成及编码技术	极化特征分离表示,散射特性保持编码	[168-169]
		特征生成及编码技术、近似重构技术	时频联合分析框架,干涉相位保持重构	[170-173]
	混合编码结构	机器视觉语义表示、近似重构技术	压缩感知增强的可解释性重构	[174-178]
高光谱	端到端编码结构	特征生成及编码技术、近似重构技术	光谱－空间解耦卷积编码,波段自适应量化	[179-184]
	混合编码结构	机器视觉语义表示、近似重构技术	边缘感知渐进重构,低秩张量分解重构	[185-186]
	集成编码结构	特征生成及编码技术、近似重构技术	联合卷积稀疏编码,波段相关性特征蒸馏	[25,187-189]

4.1　基于机器语义的视觉信息表示方法

在机器视觉领域,视觉信息表示方法的研究正朝着基于机器语义的方向发展。机器视觉的数据源种类繁多,包括可见光视频、红外图像、深度图等,其基础建模往往源于对数据源及其内在语义的抽象与分析。机器语义编码属于信息处理范畴,主要通过对视觉信息进行预处理、特征提取以及层次化语义解析,将原始数据转化为紧凑且具备内在结构的概

念表示。通过按照数据源、语义等维度对视觉信息进行分类编码，可以有效找出其基本的结构特征，为智能视觉信息处理提供理论支持。本文以可见光视频数据的表示方法为例，详细描述了其建模过程和编码策略。

现有研究表明，机器视觉语义表示能够将视觉数据抽象为紧凑的概念数据，在压缩码率的同时保留对视觉内容的清晰理解，并提供更灵活的控制[49]。通过选择适合的表示方法，在极低码率的人脸重构应用场景下，可以实现近千倍的压缩比，同时保持较高的视觉重建质量和对视觉分析任务的适应性[50]。此外，基于机器语义的表示方法已逐渐超越传统信号级处理模式，转向对高级视觉概念的抽象和处理，从而为未来智能视觉系统的发展带来了全新的机遇。

图 4 所示的机器视觉语义表示模块通过将输入的数字信号 X 投影到由基函数 $f_1, f_2, \cdots, f_n$ 组成的空间上，得到任务相关的信号表示 f^*。公式（7）说明了在任务相关信息 t 的条件下，使用函数 U 确定最适合的表示方法 f^*。通过该流程，最优数据表示方法 f^* 被确定，并使用这一方法将输入信号 X 转换为输出信号表示 Y，即 $Y=f^*(X)$。通过这种方法，不仅实现对视觉信号的紧凑表示，还能够满足任务对信号解码的具体要求，从而提升编码效率和重建质量。

$$f^* = U(f_1, f_2, \cdots, f_n X, t) \tag{7}$$

根据不同的任务需求，机器视觉的表示方法对于高效的数据处理、视觉任务的精确执行以及语义编码有重要影响。主要表示方法如图 5，包括：

（1）深度结构与纹理表示：在编码后的重建任务中，深度结构和纹理表示起到了重要作用。通过语义分割模型提取结构化边缘层和语义纹理层，可以对细粒度的纹理分布进行建模，从而在图像重建过程中提升细节构建的灵活性[51-52]。分层融合生成对抗网络 GAN（Generative Adversarial Networks）[53]利用这些比特流将纹理渲染到解码后的结构表示中，从而实现高质量和逼真的图像重建。

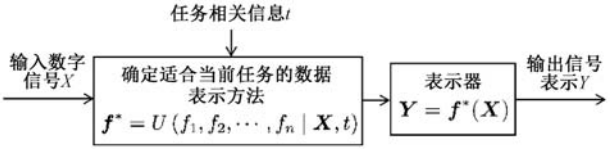


图 4 机器视觉语义表示

（2）感兴趣区域表示：在目标检测和实例分割任务中，提取与任务相关的目标区域至关重要。语义 RoI 对齐技术（Semantic RoI Align）[54-55]通过计算每个像素或区域的显著性得分，使用阈值筛选出相关区域，以便对感兴趣区域进行高效表示[56]。在目标检测、语义分割等场景中尤为关键，能准确定位和分割目标物体[57]。

（3）关键点与动态表示：关键点表示常用于物体运动编码，如在姿态估计任务中，神经网络能够分解和提取与身份及运动相关的特定信息[58]。而在视频处理中，光流表示通过描述像素随时间变化的向量场，捕捉视频帧间的动态信息，从而有效编码连续运动特征[59]。这种动态表示方法能够在视频理解和动作识别中发挥重要作用[60]。

（4）机器学习特征：基于流形学习框架，对高维特征协方差矩阵进行谱分解，选取最大特征值对应的正交基构建低维嵌入空间。通过鉴别性损失函数优化投影矩阵，确保降维后特征保留类别可分性[61]。

（5）分割图表示：对于视频理解等任务，当前的方法通常通过语义分割模型对视频帧进行处理，并

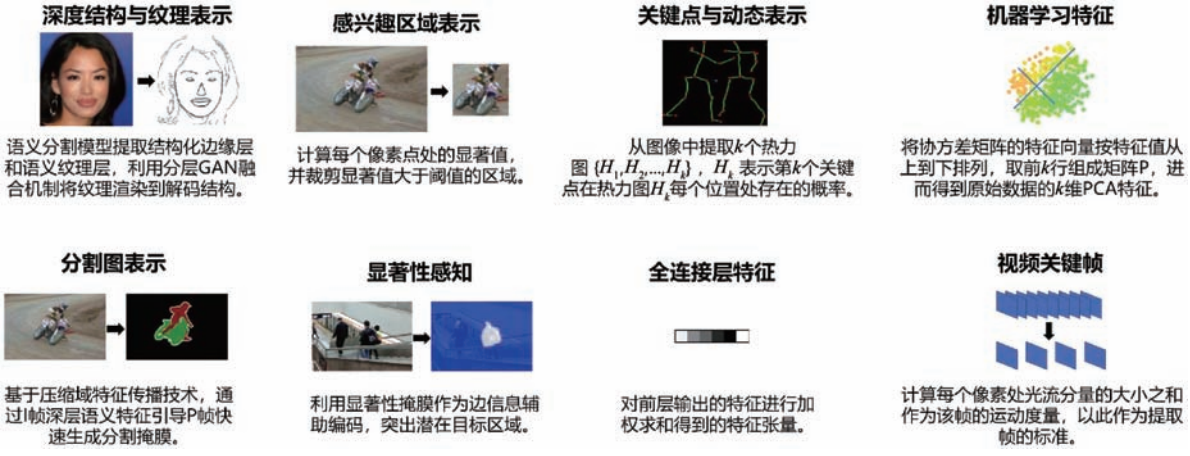


图 5 基于机器语义的多种视觉信息表示方法示意

结合压缩域的特征传播技术来加速分割过程[62]。例如,基于时序特征传播框架,利用关键帧语义信息引导后续帧快速生成分割掩膜。动态补偿模块缓解运动形变导致的特征偏移,保持跨帧分割一致性[63]。

(6)显著性感知:通过多尺度特征融合生成显著性图,结合门控机制筛选关键区域,指导编码器优化比特分配。该机制可自适应不同任务需求,在监控视频编码中优先保留人脸、车牌等高价值目标信息[64-65]。

(7)全连接层特征:对卷积特征进行全局上下文建模,通过注意力加权强化判别性特征表示,形成紧凑的高层语义表示。此类特征在跨模态检索任务中展现出强泛化能力,支持图文匹配与语义推理[66]。

(8)视频关键帧:通过运动能量分析与场景变化检测自适应选择关键帧,结合差分编码策略减少时序冗余。该框架在直播流媒体场景中平衡了压缩效率与视觉连续性需求[67]。

机器视觉中的表示方法多样化且高度依赖于任务需求和场景[68]。在编码端和解码端选择合适的表示方法对于提高数据处理效率和任务执行精度至关重要。特别是在语义推理、目标检测和视频理解等复杂任务中,合理优化表示方法将显著提升视觉系统的整体性能。

4.2　特征生成及编码技术

对用于机器任务的视觉数据进行深入分析和理解,核心在于通过特征提取来表征数据,而不是单纯的纹理识别。这些提取出的特征构成了视觉数据的高效表示,从而实现了信息的高效传递。通过对特征生成和优化流程的梳理,可以进一步理解如何从输入信号中获取有助于后续处理的关键信息[69]。

在特征生成的过程中,首先从输入信号 Y 中提取出初始特征 $F_0=g(Y)$。然而,单纯的初始特征可能无法满足任务的所有需求,因此需要通过特征决策优化来进一步提高任务性能。该优化过程的目标是,在满足特定计算成本 $C(F)\leqslant c$ 的前提下,最大化任务质量函数 $Q=\sum_{i=0}^{n}q_i(F)$。

优化后的特征 F 不仅能够更好地表征视觉数据,还能够在不增加过多计算负担的情况下提升处理效果。与此同时,特征优化过程中存在一个反馈机制,能够将优化结果传递回特征生成模块,用于调整和改进特征生成策略,确保系统的持续优化。整个优化流程及目标如图 6 所示。

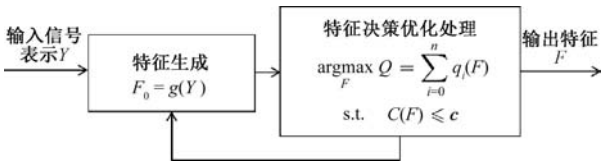


图 6　特征生成模块优化流程

特征提取编码技术不仅支持全分辨率编码,有效消除阻塞伪影,同时在广泛区域内获取更多上下文信息。近年来,基于 Transformer 的编码模型得到了广泛关注。例如,ViT(Vision Transformer)[70]通过重构图像分类流程,从压缩特征中直接进行分类,优化了图像压缩与特征提取的结合,提高了长程信息的利用效率。逐层 Transformer[71]在多用户语义通信系统中应用,通过提取文本与图像信息,融合多模态数据,增强了特征传输的效率,提升了任务执行的性能。这种方法为特征编码的灵活性提供了新的视角。此外,SimCLR 和 BYOL[72]通过引入信息压缩,显著提高了视觉表征的鲁棒性,优化了下游任务的表现。这些算法通过引入信息压缩机制,增强了特征表示的鲁棒性和泛化能力。Swin Transformer[73]利用层次化结构和移动窗口机制,优化了计算复杂度,适用于各种视觉任务,为特征生成与编码提供了高效的解决方案。EfficientNetV2[74]通过联合优化训练速度和参数效率,显著提升了模型在特征提取与编码方面的性能,满足了快速处理的需求。InceptionNeXt[75]则通过将大核深度卷积分解为多个并行分支,提升了特征提取的效率,同时保持了良好的性能,为编码技术提供了新的思路。MF-Net[76]通过多尺度特征提取与融合,优化了无监督变形配准任务,增强了图像细节的注册精度,为特征生成和编码提供了有力支持。这些先进的编码技术和模型设计不断推动机器视觉特征生成的研究与应用,显著提升了图像压缩和分析的性能。

4.3　面向任务的编码技术

在机器视觉领域,面向任务的编码技术是提高编码效率和任务准确性的重要手段[77-78]。以下将对单任务、多任务以及任务关注的特定区域编码技术进行对比和分析,并探讨其在实际应用中的优势和局限性。

(1)单任务编码技术:通常针对特定的机器视觉任务,如分类、对象检测或语义分割任务。通过预训练的 RPN(Region Proposal Networks)和 VGG 模型,使得 VVC 可以支持这些单一任务[79]。这种方法的优势在于优化了针对特定任务的压缩内容,提

高了任务执行的效率和准确性。例如,在目标检测中,使用由 Mask-RCNN 生成的语义重要性图,将更多比特分配给语义上重要的区域,从而提高目标检测和语义分割的性能[80]。然而,单任务编码的局限性在于其适应性较差。当需要支持多种任务时,单任务编码往往难以兼顾不同任务的需求。例如,针对监控场景的前景-背景分离视频压缩技术,仅在背景静止时有效[81]。

(2)多任务编码技术:旨在同时支持多个机器视觉任务,通过共享编码资源来提高整体效率。InterDigital 的 AI 实验室提出了一种"连接器"方法,该方法在推理时将主任务的重建图像转换以适应次级任务,从而在不提取多任务特征的情况下支持多个任务[82]。这种方法适用于多种应用场景,但在下游任务与主任务关系不紧密时效果有限。所以,在此基础上引入任务特定损失和特征压缩损失,训练多任务模型,实现深层次的中间特征有损压缩,而不牺牲任务准确性。虽然这种方法在传输单个深度特征张量时表现良好,但对于如 Mask-RCNN 和 YOLO 等[83-85]复杂的多任务网络效果不尽理想。

为了应对这一挑战,进而衍生出了比特分配方法用于传输多个特征流,从而更有效地支持多任务处理[86]。此外,通过基于学习的深度特征编解码器和 codebook 超先验模型,可以在低比特率下构建紧凑且富有表现力的表示,以支持多样化的机器视觉任务[87]。

(3)任务关注的特定区域编码技术:通过优化特定区域的编码资源分配来提高视觉任务的准确性和效率。使用 RPN 生成高质量建议,并结合 FPN(Feature Pyramid Network)和 Mask R-CNN 识别不同大小的 RoI 区域。这些区域经过编码资源的强化,特征表示得以提升,检测精度显著提高[88]。对于非重要区域,则采用高效编码策略,如提高量化参数 QP(Quantizer Parameter),以减少资源占用。

具体的提取过程如下:处理 $n+1$ 帧视频时,首先将其划分为一系列帧 $\{I^0, I^1, \cdots, I^n\}$。每帧 I^k 通过 FPN 等神经网络提取多尺度特征图,识别不同大小的物体并提取相关区域。如果某帧 k 包含多个对象 m,则生成一系列预测边界框,某个边界框表示为 $A^k_{m,obj}$,并通过目标存在概率图确定候选区域 $A^k_{m,obn}$。神经网络预测每层中对象的存在分数,确定检测结果来源的特征层。由于视频中相邻帧的目标通常处于相似位置,因此按图像组 GOP(Group of Pictures)对目标进行分组,以保持时间一致性。如果帧间对象的 ID 相同且交并比 IoU(Intersection of Union)大于阈值,这些区域将被合并。

$$A^k_{m,MAR} = A^k_{m,obj} \cup A^k_{m,obn} \tag{8}$$

在编码过程中,VVC 以 CU(Coding Unit)为单位进行压缩,并在同一 CTU(Coding Tree Unit)内进行空间预测。为了保持视频的一致性,需要将区域扩展到 CTU 范围。对象检测网络与压缩过程密切相关,高预测置信度通常表明对象在边界框内。然而,在高 QP 下,重叠对象可能因压缩失去特定特征而难以检测。

4.4 可变码率优化技术

可变码率优化技术 RDO 通过动态调整比特率以适应不同场景需求,提高编码效率和机器视觉任务性能。由于当前基于神经网络的端到端视频编码模型大多采用固定码率编码,为此提出了基于变分自编码器 VAE(Variational Auto-Encoders)的信号依赖编码优化模型,详细分析量化、深度特征的比特分配及失真之间的关系,从而实现可变码率的编码[89]。将 RDO 的部分使用神经网络替换,研究者提出符合标准的基于特征的 FRDO(Feature-Based Rate-Distortion Optimization)方法,通过在神经网络第一层创建的特征空间中计算失真指标替换传统 RDO 中基于像素的失真指标,从而提高机器视频编码场景中的编码性能[90]。此外,这种方法通过结合 CU 特征和部分 ResNet 架构,能够在高压缩率情况下保证机器视觉任务的性能[91]。将 ROD 完全使用神经网络实现,研究人员开发了一种速率失真优化学习分层双向视频压缩框架,该框架结合了传统分层双向运动补偿的优点和数据驱动的端到端速率失真优化[92]。此外,基于神经网络的速率控制系统能够在给定比特率的情况下对视频进行精确编码,同时增强速率失真性能[93]。

4.5 近似重构技术

该重构技术的目标是在解码阶段从压缩数据中近似重构出高质量的视觉信息。由于编码过程中不可逆的量化操作必然引入信息丢失,解码阶段只能实现近似重构,从而形成一种不可恢复的重构方式。编码阶段,原始视频序列 X 经过一系列处理步骤,包括变换编码、量化和熵编码,最终生成压缩比特流 B。这些操作旨在有效压缩视频数据,同时在一定程度上保留视觉质量,但量化过程使得部分细节不可恢复。

编码流程:

1)变换编码:将输入的原始视频序列 X 转化为频域或其他领域的系数矩阵,以减少数据冗余。

2)量化:对变换后的系数进行离散化处理,降低数据精度,从而大幅压缩数据量,但同时引入不可逆的信息损失。

3)熵编码:利用无损编码技术将量化后的系数进一步压缩为比特流 B,以去除数据中的统计冗余。

解码与重构流程:

1)熵解码:首先对比特流 B 进行熵解码,恢复出经过量化处理后的系数矩阵,此步骤仅还原编码时压缩的频域数据。

2)反量化:通过反量化过程,将量化后的系数恢复至近似原始的变换系数矩阵,但由于量化不可逆,细节信息已部分丢失。

3)反变换:对反量化得到的系数矩阵进行反变换,将频域系数重新映射为空间域图像数据,形成近似重构的图像帧。

在帧间编码的场景下,为了减少帧间冗余并部分补偿量化引起的误差,视频重构还引入了运动补偿。运动补偿利用前一帧的重构图像 $\hat{X}_{t-1}$ 及当前帧的运动矢量 M_t 预测当前帧,并修正预测误差,流程如下:

运动补偿(适用于帧间编码):

$$\hat{X}_t = \hat{X}_{t-1} + MotionCompensation(M_t, \hat{X}_{t-1}) \tag{9}$$

运动补偿通过当前帧的运动矢量 M_t 对前一帧图像 $\hat{X}_{t-1}$ 进行位移和调整,从而生成一个初步的估计图像帧。

重构过程可表示为:

$$\hat{X}_t = Reconstruction(B, M_t, \hat{X}_{t-1}) \tag{10}$$

最终的重构图像帧 $\hat{X}_t$ 是通过综合比特流 B、运动矢量 M_t 和前一帧的重构结果 $\backslash hatX_{t-1}$ 生成的。该过程不仅涉及图像数据的恢复,还包含帧间预测与运动补偿技术,以在不可恢复的重构条件下尽可能提升视觉质量和时空连续性。

近年来,深度神经网络在编码与重构领域取得了显著进展,尤其在极低比特率下生成高质量重构图像方面表现突出。有研究者提出引入创新的编码器控制机制[94],可在不同任务(如检测和跟踪)中灵活调整编码参数,通过模式预测和组帧选择模块,使预训练的深度视频压缩解码器适应多任务场景,而无需为每个任务训练独立的解码器。此外,卷积神经网络(CNN)与长短期记忆网络(LSTM)等技术的结合进一步提高了压缩效率[95]。在人脸编辑任务中,MaskFaceGAN[96]采用局部属性编辑技术,优化了预训练生成对抗网络(如 StyleGAN2)的潜在码,实现了对目标面部属性的精确编辑,同时保留了其他相关信息。未来,基于感知相似性的压缩与重构方法有望进一步提升视觉数据的压缩效率和重构质量,尽可能缓解不可恢复编码带来的信息丢失问题。

5　机器视觉的编码结构

一般来说,一个好的机器视觉编码应该具有以下特点[154]:(1)动态压缩范围大,可以根据场景复杂度动态调整压缩率,节省传输带宽;(2)高压缩率;(3)码率自适应;(4)可扩展性,在不重新编码原始信息的情况下,动态添加额外的信息,满足解码器对更高质量或更多信息的需求;(5)达到数据压缩导致的最大可接受失真程度,在此程度之上,机器视觉模型的性能将下降至不可接受的水平;(6)低复杂度,复杂度与码长呈线性关系。

目前,代表性的编码结构研究包括几种类型。基于端到端的机器视觉编码结构通过自动化简化流程,消除部分中间步骤,为执行机器任务设计;混合编码结构融合了帧间预测、变换编码和深度学习技术,提高编码效率;人机混合编码优化了码流结构,兼顾机器任务和高质量视频;一些研究结合混合编码的高效和端到端编码的自动化,提升性能;采用高级视觉特征对内容信息进行抽象表示的编码结构,从而实现了语义感知编码;点云编码采用几何和属性分离压缩,利用空间相关性消除冗余。

5.1　基于端到端的机器视觉编码结构

自 2015 年,基于端到端机器学习的视觉编码技术逐渐兴起。通过神经网络的分层结构,非线性映射得以精确建模,这不仅优化了预测模型,还显著提升了信号重构的质量。随后出现了参数化非线性变换方法,有效减少了转换分量间的互信息,为图像压缩领域带来了显著进步。2019 年,研究者们提出了结合传统视频压缩方法与神经网络强大非线性表示能力的端到端视频压缩深度模型架构[190]。这一架构对视频压缩的所有组件进行了联合优化,进一步提升了压缩效率。此外,无监督特征学习模型通过预测缺失帧或推断未来帧,发现了空间和时间相关性,这对于表示复杂的变形和运动模式具有重大意义[141]。基于变分自动编码器的端到端可训练图像压缩模型引入了超先验的概念,以有效捕获潜在表

表 2 机器视觉编码主要算法分析

数据	文献	特 点	不 足
可见光	[1]	基于码本的超先验模型，通过减少像素和神经网络特征之间的维度差距提高多任务特征的压缩效率	算法所依赖的参数量较多
	[13]	利用轻量化递归残差网络与多注意力融合，实现单模型适配多编码参数，降低参数和训练成本	增加计算复杂度和延迟
	[15]	将非局部模型与 Intra-Inter In-Loop Filtering 相结合的屏幕内容感知视频编码，提升编码效率和帧质量	计算复杂度较高，对硬件资源和能耗要求较高
	[18]	提出一种针对机器视觉的端到端图像压缩优化方案，结合了专用的可变码率优化技术	适用于特定的深度学习模型，未考虑众多新兴的目标检测方法
	[34]	训练图片最好的量化参数选择模型。在该量化参数下，图像能够在任务精度和码率上达到平衡	算法复杂度高，不适合高并发场景
	[36]	平衡损失函数，如视觉任务损失、图像失真损失和码率	图像质量依赖模型训练，不能做到软件可定义
	[50]	基于语义先验的深度表示和跨通道熵模型提升了图像压缩率和重建质量	需提前确定特定的应用场景、解码器复杂度较高
	[53]	提出一种双层语义压缩框架，将结构与纹理分开编码，提升重建质量并支持灵活的内容操作	应用场景局限，并仅适用图片数据集
	[58]	提出一种人物视频合成模型，使用低带宽生成高质量视频，并支持头部旋转，提升视频会议体验	应用场景局限
	[59]	语义结构化视频编码通过编码静态对象和动态运动信息，自适应支持多个智能任务	对复杂非线性运动处理能力有限
	[62]	提出一种统一框架，能够高效地联合压缩多种视觉和语义数据，充分利用它们之间的冗余	对于特定数据类型的适应性仍有提升空间
	[64]	提出一种新的无损压缩框架，利用自适应预测和图上下文卷积来高效存储和传输场景图数据	复杂场景图的处理和通用性需要进一步验证
	[66]	无需特殊架构或内存库的简化对比学习框架，通过优化数据增强、非线性变换等显著提升视觉表征	对编码的集成与优化仍待深入研究
	[69]	将像素信息转化为自然语言，后用自然语言完成图像在语义层面的重建	语义上能较好重建，但无法做到近似的像素重建
	[70]	提出一种基于 Transformer 的端到端图像压缩与分析模型，能够在压缩特征的基础上进行图像分类	算法所依赖的参数量较多
	[79]	通过提取机器视觉中 RoI，并优化位分配来降低比特率，同时维持高机器视觉性能	任务性能受到 RoI 预测限制
	[80]	基于强化学习(RL)的语义位分配实现了任务驱动的语义编码	任务性能受到语义预测限制
	[87]	端到端的图像压缩，并带有任务精度损失	算法复杂度高
	[91]	提出一种用于机器视觉编码的快速算法，结合 ResNet 特征提取，提升视觉任务性能并加速编码过程	解码器复杂度较高
	[94]	提出创新的编码器控制方法，使深度视频压缩能够适应多种机器任务，同时预训练解码器的兼容性	复杂视频场景的适应性和效率需要进一步验证
	[97]	设计一个信息过滤(IF)模块，在编码之前智能地丢弃冗余信息进行分析，支持多任务	任务网络仍然受到骨干网络限制
	[115]	提取特征图后使用 VVC 进行压缩，并输入视频理解任务	效果较差，部分视频与基线持平甚至有所下降
	[118]	结合 E2E 学习编解码器与传统视频编解码器优势的机器混合编解码器，提升人机编码性能，编码比 VVC 快 2 到 10 倍	解码器复杂度较高，解码比 VVC 慢 17 到 38 倍
	[125]	数字视网膜框架通过视频流、特征流和模型流联合优化提升数据压缩和任务性能	仅适合类似城市体量的大型应用场景
	[126]	提出了语义到信号的可扩展压缩方法，兼顾机器视觉和人类视觉的需求	机器任务简单
	[127]	端到端的图像压缩，并兼顾人机损失	算法复杂度高
	[128]	提出适用于人类和机器感知的压缩格式，通过学习表示的同时优化压缩效率和核心视觉任务的性能，且模型可直接从压缩数据中进行训练	当压缩率达到十倍以上，目标分割和检测任务性能会下降
	[129]	提出一种针对机器视觉任务的图像压缩预处理方法，通过神经预处理模块保留语义信息、抑制无关信息，兼顾标准兼容性、适应不同压缩比并优化任务表现	依赖传统非差分编解码器
点云	[155]	提出基于八叉树和深度学习的两阶段框架，通过利用局部体素上下文，提高点云的压缩效率	算法复杂度高，编码时间较长
	[156]	提出密度保持的深度点云压缩方法，通过自编码器架构，实现了更优的率失真权衡，保留局部密度	编码时间较长
	[158]	提出一个分层注意力结构的高效熵模型，显著提升大规模压缩性能并保持高重建效果	算法复杂度高，编码时间和解码时间较长

（续表）

数据	文献	特　点	不　足
点云	[160]	提出基于定量参数级联(QPC)和率失真优化(RDO)的算法,用于近无损点云压缩	在稀疏点云上的优化效果不显著
	[166]	提出基于 p-Laplacian 嵌入的图字典学习框架,用于 3D 点云属性的压缩	计算复杂度较高,尤其是在大规模点云数据集上
SAR	[168]	提出一种结合离散化的高斯自适应模型和广义减法归一化的 SAR 图像压缩方法	在使用 MS-SSIM 优化时,性能略低
	[169]	提出一种自监督学习的 SAR 图像压缩和去斑点联合处理方法,有效减少图像噪声并提高压缩效率	算法复杂度高
	[170]	提出结合多重残差块和局部－全局上下文熵模型的 SAR 图像压缩算法	解码器复杂度较高,解码时间较长
	[171]	提出一种基于深度学习的原始 SAR 数据压缩方法,利用自回归熵模型在较低比特率下保持了高图像质量和相位保真度	编码后的图像在低反射区域(如水体)出现较明显的噪声
	[172]	基于变分自编码器和卷积神经网络的 SAR 图像压缩方法,通过联合变换和残差块提升压缩效果	算法复杂度高
高光谱	[25]	提出边缘引导的高光谱图像压缩模型,通过互动双注意力和边缘引导损失提高结构信息的保留能力	编码时间较长
	[179]	提出一种基于卷积自编码器的光谱信号压缩网络	在高压缩比下可能会出现压缩伪影和精度下降
	[184]	提出一种基于跨通道对比学习的高光谱压缩网络,有效解决了高压缩比下信息丢失和特征塌陷问题	处理复杂高光谱数据时,特征表示不足,导致压缩数据与下游任务存在语义差距

示中的空间依赖关系[133]。在保持端到端优化的同时,研究者还探索了自回归、分层和组合先验等替代方案,并在图像压缩背景下权衡了它们的成本和收益[134]。这些研究不仅推动了视觉编码领域的发展,也为未来的图像和视频压缩技术提供了新的思路和方向。

面向机器视觉的端到端编码结构如图 7 所示,包含以下步骤:

(1)输入视觉信号:首先,将原始视觉信号 X 输入到编码器中。

(2)编码:编码器将视觉信号转换为新的数据表示形式 y,即 $y=E(X;\theta_E)$,其中 θ_E 代表编码器的参数。

(3)量化:编码后的数据 y 进一步被量化为潜在码 z,表示为 $z=Q(y)$。

(4)熵编码:通过算术编码器,利用熵模型估计的概率分布对潜在码 z 进行无损压缩,生成紧凑的比特流。同时,根据熵模型计算紧凑的代价 R。

(5)比特流传输:压缩后的比特流通过传输信道发送到解码端。

(6)解码潜在码:在解码器端,算术解码器对传输的比特流进行解码,恢复潜在码 $\hat{z}$。

(7)重建视觉数据:经过重建过程,潜在码 $\hat{z}$ 被转换为重建的视觉数据 $\hat{X}$,即 $\hat{X}=D(\hat{z};\theta_D)$,其中 θ_D 代表解码器的参数。

(8)输出视觉数据:最终,解码器输出重建后的视觉数据 $\hat{X}$,可供后续的机器视觉任务算法处理。

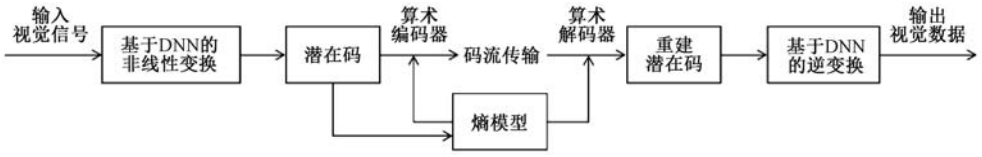


图 7　端到端典型编码结构

通过上述步骤,端到端编码结构能够实现高效的视频压缩与解码,从而为机器视觉任务提供高质量的输入数据。视频编码的目的不仅仅是传输或存储视觉数据,更在于支持机器视觉任务的执行。在实际应用中,特别是涉及复杂环境中的机器视觉时,系统往往需要在压缩视觉数据的同时,保留足够的图像信息以供后续分析。

先分析后编码:通常,机器视觉依赖于视觉特征进行图像的分析与理解。为了提高分析效率,前端设备首先提取这些视觉特征,然后将其编码并传输至后端设备,如负责机器智能处理的云端服务器。这一流程通常被称为“先分析后编码”,如图 8 所示。通过在前端完成特征提取与分析,大大减少了数据传输的复杂性[125]。同时,以紧凑的方式表示和传输这些视觉特征,不仅降低了视觉数据表示的成本,还极大地满足了大规模视觉分析应用对带宽消耗的严苛要求。例如,结合可变码率编码和广义速率精度优化,共同优化了压缩和机器视觉网络,充分利用

了机器视觉对编码图像的最大潜力。此外，通过学习预测缺失帧或从输入视频序列中推断未来帧，这种方案发现了空间和时间相关性，有效模拟实际应用中的编码不确定性[105]。

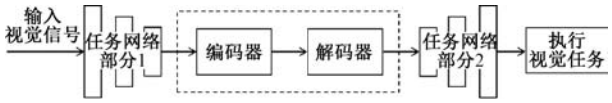


图 8 “先分析后编码”结构

与“先分析后编码”不同，另一种处理流程是“先编码后分析”，如图 9 所示。前端设备首先对原始视觉数据进行压缩，然后将编码后的数据传输到后端设备。后端设备在接收到数据后，再进行解码和分析，提取所需的视觉特征。这种方式在前端大大减少了算力需求和预处理时延。然而，它也可能导致压缩过程中信息损失，影响后端设备对视觉特征的准确提取和分析[115]。因此，在采用“先压缩后分析”时，需要权衡压缩比率和信息保留度，确保降低传输成本的同时，不影响后端设备的分析精度[109]。为了克服这一限制，深度上下文视频压缩框架通过传播特征来学习多尺度的时间上下文，并将学习到的时间上下文重新填充到压缩方案的各个模块中，从而实现从预测编码到条件编码的范式转变[110]。

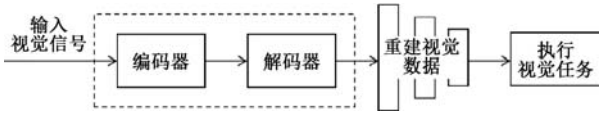


图 9 “先编码后分析”结构

5.2 混合编码结构

在机器视觉编码中，传统面向人类视觉的编码器可被重用，这为采用成熟的编码技术处理机器视觉任务提供了有效途径。在机器视觉的混合编码结构中[191]，编码过程采用基于块的逻辑，将图像划分为若干编码单元，并利用内部和间预测减少视频数据的空间和时间冗余。根据图 10 所示，各符号和箭头指示了视频编码器的工作流程。实线箭头表示数据流的处理顺序，从输入的视频信号开始，经过预测、残差计算、变换、量化等步骤，最终输出压缩码流。符号⊕表示在预测和残差重建过程中进行的加减运算。具体的编码过程可详细分为以下步骤：

(1)宏块划分：首先，图像被划分为宏块 MB(Macroblock)，并进一步分割为多个切片，每个切片独立编码。每个切片又被划分为编码树单元 CTU，CTU 中包含编码单元 CU、预测单元 PU(Prediction Unit)和变换单元 TU(Transform Unit)。这

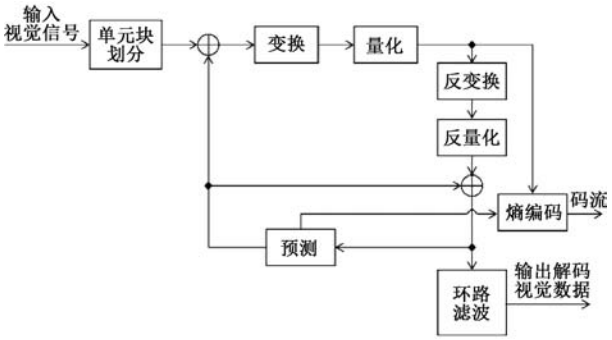


图 10 混合编码结构

确保了编码的灵活性，允许对不同部分图像使用不同的编码、预测和变换策略。

(2)预测阶段：预测阶段分为帧内预测和帧间预测[91,116]。帧内预测利用当前 PU 上方和左侧的先前重建像素作为参考，通过变换单元生成预测值。帧间预测则通过运动补偿处理图像块，假设块内的运动均匀，解决移动物体跨越多个块的情况。通过对先前帧中的参考块进行搜索，能够减小视频中的时间冗余。

(3)残差计算：对原始图像块与预测图像块进行相减，生成残差数据块。这一步骤的重要性在于减少冗余信息，并为后续的变换和量化做好准备。

(4)变换与量化：残差数据块经过变换和量化，进一步压缩信息，将能量集中于较少的变换系数中。这使得数据在后续的熵编码中能够以较少的比特数表示。

(5)反变换与反量化：为了重建图像，量化后的系数数据经过反量化和反变换，得到变换后的残差数据块。

(6)重建图像块：将变换后的残差数据块与预测图像块相加，生成重建图像块。这个重建图像不仅用于当前帧的显示，还会作为后续帧预测的参考[120]。

(7)环路滤波：由于块状处理方式，重建图像块可能产生块效应(block artifacts)。通过环路滤波器(Loop Filter)，能够平滑处理块边界，消除伪影并提高视觉质量。

(8)熵编码：最终，经过量化和预测的系数数据与预测模式一起进行熵编码，生成压缩后的码流供输出。整个编码过程通过速率失真优化，平衡编码效率与图像质量。

通过以上步骤，图像逐块被压缩为码流，混合编码框架在编码过程中利用多种工具进行优化，以最大限度减少数据冗余并提高压缩效率[192]。

在传统混合视频编码框架中,多个模块可用深度神经网络替换,包括 RDO、内预测、间预测、量化、熵编码和滤波器[107-108,121],以提升整体性能和视觉质量。针对每个模块的缺点,研究者们提出了多种创新解决方案。

首先,针对 RDO 模块忽略人类视觉系统 HVS(Human Visual System)特性的问题,研究者开发了基于学习的客观评估指标 VMAF(Video Multi-method Assessment Fusion),并将其纳入 RDO 中,从而提高感知编码效率和视觉质量[104]。此外,提出了速率感知深度预处理的概念,即在每个输入帧上进行一次传递,优化视觉质量,使得在任何比特率和编解码器压缩视频时,都能显著提高其视觉质量[122]。传统 RDO 广泛使用均方误差进行速率失真优化,虽然能提供较高的峰值信噪比,但感知质量欠佳。为此,研究者从端到端图像压缩中提炼"感知"知识,用以增强多功能视频编码 VVC 内部编码的感知质量[123]。为了无缝适应时变网络带宽,研究者开发了质量可扩展高效视频编码,并提出了一种快速决策算法,降低质量可分级编码的复杂度,同时保持高效的视频传输和处理性能[124]。

5.3　人机混合编码结构

为了缩小机器视觉编码的语义特征与人类视觉编码的像素表示之间的差距,研究者们构建了一种高效且可扩展的人机混合编码结构[135]。通过优化多个特征流之间的位效用[193],旨在联合提升人眼感知和各种机器视觉任务的性能和编码效率。此框架整合反馈机制,实现视频、特征、模型等多流的协同编码,满足人类感知与机器智能的多任务需求,如图 11 所示。

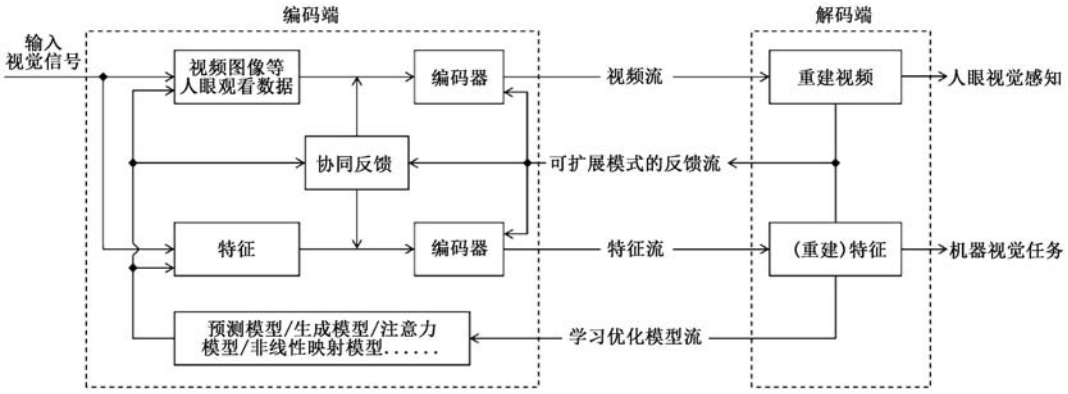


图 11　人机混合编码结构

编码端:

(1)输入视觉信号:系统接收到的视觉信号 X 进入编码端,系统首先从输入信号中提取出特征集 $F=\{F^0,F^1,\cdots,F^n\}$,其中 $V=F^n$ 代表像素特征,专注于提升人眼观察者的视觉保真度。当 $0\leqslant i<n$ 时,F^i 代表与机器视觉任务相关或无关的语义、句法特征,且编号越小,特征越抽象。

(2)任务性能评估:每个任务 i 的性能 q^i 可以通过公式评估:

$$q^i=L^i(\hat{F}^i) \tag{11}$$

其中,$L^i(\cdot)$ 是任务 i 的质量评估标准,$\hat{F}^i$ 为经过编解码后的重建特征。某些压缩过程中,系统并不要求完全重建特征,因此可以根据实际需求进行任务性能的动态调节。

(3)编码与特征压缩:通过函数 $E(\cdot|\theta_e)$ 进行编码,即对输入特征 F 进行压缩。不同的任务可能会产生不同的编码需求,并且通过压缩后输出视频流和特征流:视频流是经过压缩的像素特征 F^n 被编码为视频流,用于人眼感知视觉保真度。

特征流是其他任务相关特征 F^i 则通过相应的编码器进行压缩,生成供机器视觉任务使用的特征流。

(4)预测模型与反馈机制:为了提升编码和压缩效率,系统利用预测模型 $P(\cdot|\theta_p)$ 对特征进行预测,通过协同反馈机制,反馈信息流实时优化系统的压缩与编码策略。

(5)资源消耗与优化目标:系统的设计目标是优化任务性能 q^i 和资源消耗 $S(\cdot)$。考虑到压缩、传输和解码的资源开销,构建了一个优化目标函数:

$$\max_{\Theta=\{\theta_e,\theta_d,\theta_p\}}\sum_{0\leqslant i\leqslant n}\omega^i q^i,\ \text{s.t.}\sum_{0\leqslant i\leqslant n}\omega^i=1 \tag{12}$$

该函数约束条件为资源消耗不能超过系统总资源限制 S_T,且每个任务的权重 ω^i 用于调整任务性能在优化中的权重。

(6)资源消耗的计算：

$$S(R_{F^0})+\sum_{i>0}\min_{0\leqslant j\leqslant i}\{S(R_{F_{i\to j}})\}+S(R_M)+S(\Theta)\leqslant S_T \tag{13}$$

其中，$S(R_{F^0})$ 代表基础特征的资源消耗；$R_{F_{i\to j}}$ 表示任务 i 到任务 j 的传输资源消耗；$S(R_M)$ 代表模型资源消耗；$S(\Theta)$ 代表模型使用、更新等额外消耗。

解码端：

(1)视频流解码：解码端通过解码函数 $D(\cdot|\theta_d)$ 对接收到的压缩视频流进行重建，生成适合人眼观察的视频：

$$\hat{F}^0=D(R_{F^0}\theta_d) \tag{14}$$

(2)特征流解码：对于特征流的解码，则通过预测模型 $P(\cdot\theta_p)$ 进行辅助还原：

$$\hat{F}^i=D(R_{F_{i\to j}}\theta_d)+P(\hat{F}^j,i\theta_p),\text{for}\,i\neq 0 \tag{15}$$

其中，$\hat{F}^j$ 表示解码后的中间特征，该式表明了预测模型的参与可以提升特征的解码质量。

(3)反馈流优化：解码端生成的反馈流会被传回编码端，通过协同反馈机制对编码器进行动态优化。反馈机制不仅能提升视频和特征的编码效率，还能根据任务需求实时调节编码参数。

(4)优化与重构：通过调整编码器 $E(\cdot|\theta_e)$、解码器 $D(\cdot|\theta_d)$ 及预测模型 $P(\cdot|\theta_p)$，可以重新定义整个编码—解码框架，从而达到优化视频流和特征流的编码效率，并在不同的任务中均衡人眼视觉保真度与机器视觉任务性能。

该人机混合编码结构通过结合多种任务优化技术，确保在有限的资源条件下实现最佳的视频质量和特征流传输效果。这种方法对提升视觉信号的编码效率以及实现高效的人机协作具有显著的作用。

5.4 集成编码结构

在机器视觉领域存在一种集成编码结构，这种结构融合了混合编码、端到端编码以及人机编码的多种技术[118,133]。具体而言，系统中采用了一种自学式图像编码器 LIC(Self-supervisedly Learned Image Codec)负责帧内编码，这展现了端到端编码的简洁性和高效性。在帧间编码方面，该技术结合了 VVC 中的经典编码工具，利用 LIC 编码的帧作为参考，确保了编码的稳定性和兼容性，这体现了混合编码方法的灵活性，如图 12 所示。

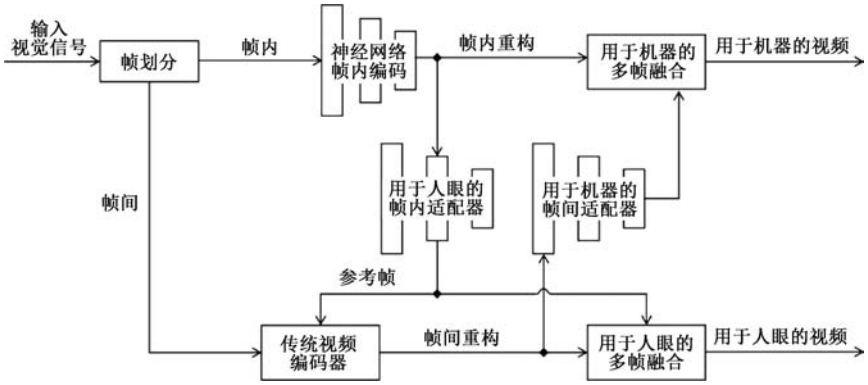


图 12 集成编码结构

在集成编码结构中，视觉信号 X 的编码过程分为多个步骤，并结合神经网络和传统的视频编码技术。这一结构不仅提高了视频压缩的效率，同时确保了针对不同任务(如机器分析和人眼观看)的视觉质量。以下是编码流程的详细说明，按步骤进行描述。

(1)输入信号分解：首先，输入的视觉信号 X 经过帧划分模块处理。在这个步骤中，视觉信号被分解为两种特征集：帧内特征 $\boldsymbol{F}^{intra}$ 和帧间特征 $\boldsymbol{F}^{inter}$。帧内特征用于在一帧中捕捉静态信息，而帧间特征则用于描述帧与帧之间的动态变化。

$$X\to(\boldsymbol{F}^{intra},\boldsymbol{F}^{inter}) \tag{16}$$

(2)帧内编码：在帧内编码阶段，帧内特征 $\boldsymbol{F}^{intra}$ 被输入到神经网络编码器中。该编码器提取出高度压缩的纹理和结构信息。通过该编码，帧内信息被转换为低维度的紧凑特征表示，以减少数据量并加速传输。

(3)帧间编码：帧间特征 $\boldsymbol{F}^{inter}$ 则经过传统的视频编码器，处理利用现有的视频压缩算法(如 H.264 或 HEVC)将帧间信息压缩为比特流以便于传输和存储。这些比特流在解码器端通过传统的解码方式恢复帧间信息，并与参考帧融合，生成完整的可观看视频。

(4)帧内重构与适配：在解码端，压缩后的帧内特征通过神经网络帧内适配器进行进一步处理。该适配器根据特定的任务需求(如机器视觉或人眼观看)对特征进行调优，使得解码后的视觉信息符合预

期质量。重构过程通过以下公式描述,帧内特征 $\hat{F}^i$ 在解码后会进行重新调整:

$$\hat{F}^i = D(R_{F_{i\to j}}\theta_d) + P(\hat{F}^j, i\theta_p) \tag{17}$$

其中,$D(\cdot)$ 是解码函数,$P(\cdot)$ 是帧特征的预测函数,结合参考帧 $\hat{F}^j$ 完成最终的帧内重构。

(5)多帧融合:经过帧内和帧间编码后的视频信息进入两个独立路径。一条路径专注生成适合机器视觉的高效视频信号,主要通过神经网络的多帧融合模块优化,用于目标检测或动作识别等任务。另一条路径经过帧间重构与适配,调整后的压缩特征通过多帧融合生成高保真度的视频内容,确保人眼观看的视频质量接近原始信号。

(6)优化目标函数:为了确保不同任务的效果最优,并在给定资源限制下最大化压缩与视觉质量,编码过程通过如公式(12)目标函数进行优化。值得注意的是,该技术还包含一种后备模式,允许在特定情况下使用 VVC 进行帧内编码,从而增加了系统的选择性和适应性。

5.5　基于内容信息抽象表示的编码结构

在机器视觉中,图像内容信息主要包括两类:结构信息和纹理信息。结构信息描述的是图像中物体的几何形状、边缘和空间关系,而纹理信息则反映了物体表面的细节、颜色以及材质特性。如图 13 所示,基于内容信息抽象表示的编码框架正是依托这两种互补的视觉组件来实现高效的压缩任务[53]。该框架首先将视觉数据分解为结构和纹理两部分,然后通过提取它们的紧凑表示,不仅确保了整体图像结构的准确还原,同时也尽可能保留细节信息,从而在压缩过程中达到效率与质量的平衡[127]。

编码过程:

(1)输入视觉信号:输入视觉信号 X 被分解为两部分——结构图和纹理表示。这是通过不同的特征提取模块来实现的。

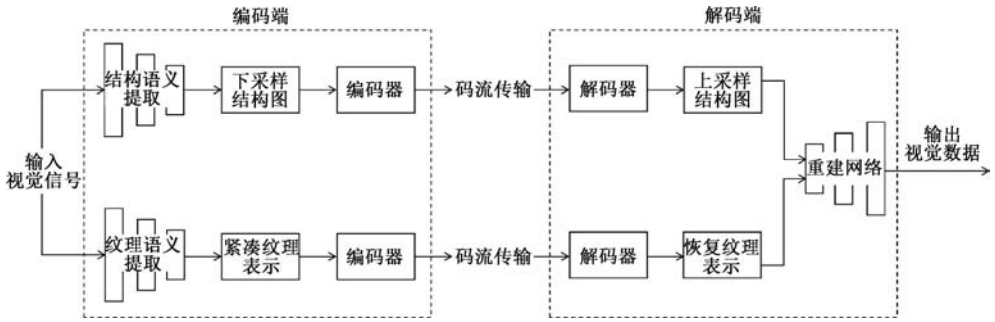


图 13　基于内容信息抽象表示的编码结构

(2)结构提取:首先,结构提取器 E^s 通过边缘检测等技术从输入视觉信号 X 中提取稀疏的结构图,表示为 $X^s = E^s(X)$。结构图 X^s 是视觉信号中的轮廓或主要的几何形态。

(3)纹理提取:与此同时,纹理提取器 E^t 使用变分自编码器 VAE 从视觉信号 X 中提取低维紧凑的纹理表示,表示为 $X^t = E^t(X)$。这一部分代表了视觉信号中的细节。

(4)编码:提取后的结构图 X^s 和纹理表示 X^t 分别进行下采样与紧凑表示,之后通过标准的编解码器进行比特流压缩,准备传输。这两个部分都被压缩为比特流,结构图 X^s 经过下采样,纹理表示 X^t 经过紧凑编码。

解码过程:

(1)解码结构图与纹理表示:在解码器端,接收到的比特流被解码为压缩的结构图 $\hat{X}^s$ 和压缩的纹理表示 $\hat{X}^t$。

(2)重建视觉信号:随后,解码得到的 $\hat{X}^s$ 和 $\hat{X}^t$ 被送入分层融合生成器 Gen,通过融合这两部分信息来重建出高质量的视觉信号 $\hat{X}$。重建过程表示为:

$$\hat{X} = Gen(\hat{X}^t, \hat{X}^s) \tag{18}$$

(3)优化与损失函数:为了确保重建的视觉信号具有高保真度,编码解码系统中引入了对抗性训练(如 GAN)和多角度失真测量技术,以监控并优化重建过程。此外,使用潜在回归损失来增强纹理的学习与重建能力。

此框架不仅提高了压缩率,还利用结构和纹理的分离再融合技术,为视觉的高品质重建与灵活的内容操作奠定了基础。同时,由于深度生成模型对统计特征的捕捉能力,该框架还展现了在纹理模式与合成范式共同学习方面的潜力,从而在跨模态上[153]实现了视觉特征与基础数据身份的统一。

5.6 点云混合编码结构

本方法以点云数据作为输入，复用了传统的点云编码结构，并在此基础上探讨了改进空间。具体来说，传统点云编码通常将视觉信息中的几何编码与属性编码分开处理。在几何编码阶段，常采用八叉树或预测树方法对点云的空间布局进行紧凑表示；而属性编码则依赖于重构的几何体，通过利用空间相关性有效消除属性数据中的冗余。尽管这种基于传统结构的编码方案在实际应用中已展现出良好的性能，但机器视觉分析的专用编码结构仍然是未来亟待突破的研究方向。

传统的几何压缩方法基于树状结构熵模型，通过利用先节点的引用以及引入与当前编码八叉树节点相同分辨率的相邻体素，实现高效的压缩。八叉树编码框架被广泛应用于几何信息的无损压缩[155]。该方法的核心在于递归地将点云划分为八叉树节点，直到达到叶节点，并通过利用相邻节点和父节点作为条件进行熵上下文建模[161]，有效地压缩了节点的占用情况。此外，为了重建更精细的物体表面，还引入三角形网格作为补充[162]。

随着深度学习技术的快速发展，基于深度学习的几何压缩方法已被证明比传统方法具有显著的编码优势[156,164]。其中，使用 3D 卷积神经网络实现基于学习的点云几何压缩成为了一种有效手段[165]。这种方法将点云视为均匀体素网格上的二进制信号，并将解码过程转换为二元分类问题。另外，特征空间预测网络也被提出[157]，它通过前一帧预测当前帧的潜在表示，并结合稀疏卷积和 3D 特征学习，以及运动补偿预测器在特征域中进行映射，最后通过概率模型压缩并传输残差。

此外，上下文信息起着关键作用。利用数百个先前解码的同级节点作为上下文，为熵编码器提供了更多的参考信息，并有效地保留了点云的详细特征[158]。为了进一步提高注意力模型的效率，提出了分层注意力和分组上下文结构[159]。这种结构不仅能够提高在大规模背景下依赖关系的探索效率，还能保持全局感受野，从而在优化压缩性能的同时，提高了编码的效率和准确性。

属性压缩涉及对点云中每个点的颜色、纹理等附加信息的编码和压缩。已有的方法分为基于传统解析变换和基于深度学习的两类框架。基于解析变换的方法依赖预定义的变换，只能表征预定义结构，如图傅里叶变换 GFT（Graph Fourier Transform）[194] 和 K-L 变换（Karhunen-Loève Transform）[195]。GFT 通过构建拉普拉斯矩阵来准确表示属性间的相关性，但由于特征分解复杂度，实时处理效率较低。为提高计算效率，提出了 p －拉普拉斯嵌入图字典学习框架[166]，利用可变信号统计和高阶几何结构进行三维点云属性压缩，建立了一种高效的交替优化范式。还有区域自适应分层变换等算法，通过自适应拟合三维点云的不规则几何结构提高压缩效率。

相比之下，基于深度学习的属性压缩能够充分利用点云数据的不规则结构，并学习到适合该数据的特定变换。例如，基于相邻层依赖关系的分析，建立了点云属性压缩的相关失真－量化和速率－量化模型，实现了 RDO 的量化参数级联算法[160]。通过将 3D 点云的属性映射到 2D 网格[167]，并利用传统 2D 图像编解码器压缩折叠图像，实现了高效的属性压缩。此外，还有基于 PointNet 的自动编码器和基于稀疏卷积的变分自编码器[196]，这些深度学习模型能更好地捕获点云数据中的复杂结构和统计特性。

图 14 是一种块自适应点云属性编码结构[163]，具体的编码流程如下：

(1)点云数据的体素化处理：首先，输入的点云数据经过体素化处理。这一步骤的目的是对三维空间中的点云数据进行离散化，将连续的点云转换为规则的网格结构。

(2)渐进聚类划分：体素化后的点云数据通过渐进聚类划分方法进行进一步的细分，将点云数据划分为多个子集或块，以便后续的编码步骤可以更有

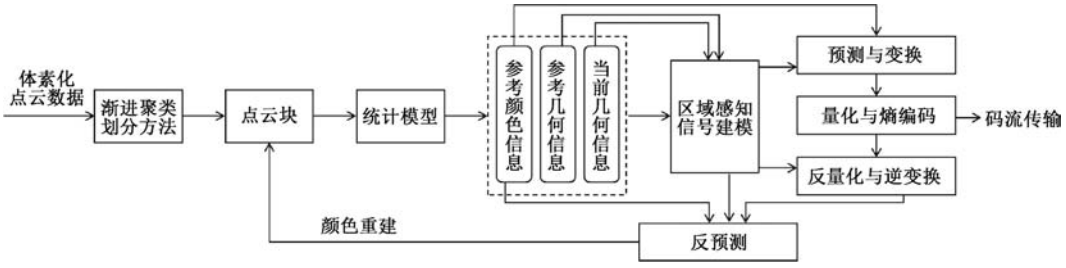


图 14 点云编码结构

效地处理这些子集。划分依据是几何相似性或颜色属性等信息。

(3)点云块的生成:划分后的点云数据被组织成点云块。这些点云块作为编码的基本单位,每个点云块包含多个点,并且这些点可能有各自的几何和属性信息。点云块将作为后续统计建模和编码操作的处理单元。

(4)建立统计模型:此模型通过分析点云块的几何信息,推断出这些点云的空间分布和属性相关性。

(5)参考颜色与几何信息的处理:在编码过程中,系统同时处理参考颜色信息和当前几何信息。参考颜色信息是为了在预测阶段提供更准确的估计,而当前几何信息是点云的基础属性。这些信息被输入到区域感知信号建模模块中。

表 3　机器视觉编码主要算法索引

数据类型	输入	数据集	算法名称	比特率 bpp↓	视觉任务	评价指标	任务精度	编码时长(秒)	参数量(M)
可见光	图片	Taskonomy[197]	codebook-hyperprior[1]	0.013	语义分割	mIoU↑	29.35%	0.107(CPU)	67.85
			[87]	0.013	分类	Top1↑	93.58%	—	—
				0.013	目标检测	mIoU↑	29.35%	—	—
	图片	MS COCO[198]	GRAO[18]	0.88	目标检测	mAP↑	62%	—	43.7
			JRD[34]	0.042	分类	Top1↑	85.80%	—	—
				0.045	目标检测	mAP↑	75.80%	—	—
			SSVC[59]	0.115	目标检测	mAP↑	39%	—	—
			[79]	0.06	目标分割	AP50↑	35%	—	—
			[97]	0.22	目标检测	mAP↑	32.20%	—	—
			[129]	0.25	目标检测	mAP↑	40.20%	0.00423(GPU)	9.42
	视频	TVD[199]	[88]	0.16	目标检测	mAP↑	53%	—	50
				0.16	目标分割	mAP↑	42.8%	—	
			[118]	0.053	目标检测	mAP↑	40%	0.09(CPU)	0.792
	图片	Cityscapes[200]	[36]	0.05	目标检测	mAP↑	20%	—	—
				0.05	目标分割	mAP↑	17.50%	—	—
			[78]	1	目标检测	mIoU↑	73.60%	—	—
			[91]	0.6	目标检测	AP50↑	62.20%	5.23(CPU)	0.34
	图片	CelebAMask-HQ[201]	[50]	0.021	面部特征点检测	NRMSE↓	43.2%	—	—
			HF-GAN[53]	0.099	面部特征点检测	NRMSE↓	40%	—	—
	图片	FGVC[202]	[126]	0.474	分类	Top1↑	88.4%	0.25(GPU)	54
	图片	TSC[203]	[80]	0.12	分类	Top5↑	81%	—	—
				0.144	目标检测	mIoU↑	98%	—	—
				0.395	目标分割	mIoU↑	97%	—	—
	图片	ImageNet[204]	[79]	0.03	分类	Top1↑	73%	—	—
	视频	ImageNetVID[204]	[127]	0.06	目标检测	mAP↑	71.90%	—	—
		MOT17[205]	[127]	0.04	目标跟踪	MOTA↑	52.6%	—	—
点云	点云数据	Semantic KITTI[206]	VoxelContext-Net[155]	0.48	重建	PSNR↑	53.19	0.109(GPU)	2.15
			[156]	1.61	重建	PSNR↑	48.5	0.08(GPU)	0.7
			EHEM[158]	1	重建	PSNR↑	62.2	2.53(GPU)	13.01
		MPEG PCC[207]	[160]	0.5	重建	PSNR↑	33.5	—	—
			p-Laplacian[166]	0.5	重建	PSNR↑	31	330.06(CPU)	—
SAR	图片	Sandia[28]	GSN[168]	0.83	重建	PSNR↑	34.571	—	6.02
			[170]	0.96	重建	PSNR↑	30.5	1(GPU)	—
			[172]	1	重建	PSNR↑	32.8	—	2.41
	图片	TerraSAR-X[208]	[169]	0.366	图像去斑	ENLs↑	266.67	—	—
高光谱	图片	Chikusei[209]	CENet[25]	0.238 (bpppb)	重建	PSNR↑	30.59	3(GPU)	35.6
			HCCNet[184]	0.228 (bpppb)	重建	PSNR↑	30.3	0.13(GPU)	41.5
	图片	VIRTIS-Rosetta[210]	SSCNet[179]	0.58 (bpppb)	重建	PSNR↑	64.85	0.116(GPU)	14.7

注:表格中"—"表示参考文献未提供相关数据。编码时长由参考文献提供。

(6)区域感知信号建模:根据参考几何信息和当前几何信息,构建出局部区域内的预测模型。目的是通过结合点云的空间和颜色信息,预测出每个点云块的最佳编码方式。

(7)预测与变换:在信号建模之后,进入预测与变换阶段。通过前面的建模结果,系统能够预测出点云块的颜色信息和几何信息,随后将这些预测结果进行数学变换,进一步减少冗余数据。

(8)量化与熵编码:经过变换的预测数据需要进行量化和熵编码。量化步骤将连续值数据转化为离散值,以便更好地压缩。而熵编码则通过进一步去除统计冗余,生成高效的比特流,以减少传输或存储的带宽要求。

6 机器视觉编码算法与应用

本章系统性地介绍了多种机器视觉数据类型的编码算法及其应用,包括可见光、点云、SAR 和高光谱数据。表 3 汇总了近三年内的主要编码算法,按照不同数据集分类,详细对比了各类算法在不同视觉任务中的表现。码率 R 通过比特率 bpp 指标体现,任务质量 Q 通过评价指标来衡量任务的精度和重构效果,而编码时长和模型参数量则反映了算法的可计算性 C。通过对这些数据和算法的对比分析,本章深入探讨了不同算法如何在 R-Q-C 之间取得平衡。特别是在典型的应用场景中,分析了各类算法的优势与不足,并给出了系统性的解决方案。

6.1 可见光数据的编码算法与应用

在可见光编码任务中,码率 R、任务质量 Q 和可计算性 C 之间存在显著的正相关关系。尽管算法不同,但整体的实验效果趋势一致。以 MS COCO 数据集[198]为例,当同算法的比特率 0.88 [18]与 0.115[59]相比,目标检测任务的 mAP 相差了 23%,这表明更高的码率能够保留更多的视觉信息,从而支持更复杂任务的精度提升。然而,任务精度的提升通常伴随着更复杂的模型结构和更高的计算需求,编码时间和模型参数量也随之增加。部分算法的模型参数量超过 50M[1,88,126],编码时长及计算复杂度显著增加。因此,如何在 R-Q-C 之间实现有效平衡,是实际应用中的关键挑战。在嵌入式设备等计算资源受限的场景中,参数量较小(<10 M)[91,118,129]、编码时长控制在 0.1 秒以内的模型能够满足部分端侧场景应用需求;而参数量较大的模型(>50 M)更适合部署在配备 GPU 计算能力的云侧服务器上,以充分发挥其精度优势。

在线上会议等高实时性需求的场景中,编码端可通过只压缩传输面部特征点,解码端生成流畅且精准的面部表情,从而在保证实时性的同时,减少数据传输量[50,53]。在工业领域,例如智能制造和自动化检测中,包含高帧率的视频处理任务,如缺陷检测和目标分类,需要极高的编码速度以满足高速流水线检测需求[69]。

在智慧城市和自动驾驶的应用中,摄像头需要长期监控复杂动态的环境,如目标检测和语义分割任务,这类场景不仅需要人类进行实时观察,还要求机器高效执行自动检测任务[128]。为应对这一需求,编码算法可以根据目标检测的优先级动态调整码率。例如,在重要目标(如车辆和行人)的检测中,采用较高码率来保留更多细节,而对背景部分则采用较低码率进行压缩[111-112]。未来如何在有限的硬件资源下维持高效的视觉处理能力将成为持续的研究热点,以加速视觉编码在各类场景中的应用。

6.2 点云数据的编码算法与应用

点云数据与可见光数据相比,在实时性和可计算性 C 方面表现较差,码率 R 和重建任务质量 Q 之间更难平衡。其原因主要是可见光数据的二维空间内的数据结构简单、体积较小,而点云涉及大量的三维几何信息和密集属性信息,导致编码码率和时延显著增加。例如,在 Semantic KITTI 数据集上,VoxelContext-Net 算法[155]在 0.48 bpp 下虽然只达到了 53.19 的 PSNR,但其编码时长短,模型参数量仅为 2.15 M,体现了较高的计算效率。相比之下,EHEM 算法[158]虽然在 1 bpp 下达到了 62.2 的 PSNR,但由于采用了分层注意力结构,计算复杂度显著增加,编码时长延长,模型参数量也增加到 13.01M。

传统的 MPEG PCC 方法与现代深度学习模型相比则表现不佳。p-Laplacian 算法[166]在 0.5 bpp 下仅能实现 31 的 PSNR,且 CPU 上编码时间长达 330.06 秒,远远落后于 GPU 加速的深度学习方法。深度学习虽然提升了压缩性能和重建质量,但其高计算复杂度和资源消耗仍然是当前的一大挑战,尤其是在高精度任务中。

点云数据编码在自动驾驶、增强现实(AR)和虚拟现实(VR)等需要处理大量三维数据的场景中有着广泛的应用。在自动驾驶领域,编码后的点云数据用于感知周围环境,以避免碰撞,对实时环境感知对数据的压缩效率提出了极高的要求。可以通过引

入大规模上下文和自回归压缩策略,实现点云几何数据的高效压缩和重建,支持高效的环境感知和实时路径规划[160]。在 AR 和 VR 的应用中,点云编码有助于提供更加沉浸式的体验,能够在有限的带宽内传输高质量的三维内容,从而实现实时的动态重建和互动。这需要算法保持点云局部密度实现高精度重建,尤其在较低比特率下能够有效保留三维场景的几何细节[157]。未来如何在保持高重建精度的同时降低计算开销,依然是点云压缩领域亟需解决的挑战。

6.3　SAR 数据的编码算法与应用

在对 SAR 视觉数据的编码研究中,目前已有多种算法展示了在降低码率 R 和高质量图像重建任务 Q 方面的显著进展。然而,编码时间尚未得到充分优化,尤其是在实时性要求高的应用中面临较大挑战。计算复杂性 C 是影响这些算法在实际场景中可行性的重要因素,目前长编码时间限制了算法的部分应用。实验数据显示,在 Sandia 数据集中,GSN 算法[168]在码率为 0.83 的情况下实现了 34.571 的 PSNR,展现了较好的图像重建质量,但其高计算成本和较长的编码时间是其一大短板。同样,局部－全局上下文熵算法[170]在 0.96 的码率下,PSNR 有所下降至 30.5,其编码时间也在使用 GPU 的情况下超过了 1 秒,反映了其计算效率的不足。

在实际应用中,SAR 视觉数据的编码技术广泛用于灾害监测、遥感成像与地形绘制等多个场景。SAR 具有全天候、全天时的优势,能够在极端天气条件下提供持续的高分辨率地表图像。通过先进的编码算法[171]减少数据传输延迟,可以在带宽受限的情况下实现大规模的数据传输,从而为应急响应提供重要保障。在城市规划和农作物监测等数据量大、传输实时性要求高的应用场景中,低码率、高质量的编码技术至关重要。通过深度学习的原始 SAR 数据压缩方法,利用自回归熵模型在较低比特率下保持了高图像质量和相位保真度,提高了数据传输的可行性和灵活性[172]。在环境监测和自然资源管理等领域,由于环境动态变化的不可预测性和监测范围的广泛性,编码算法需在低码率下实现高质量图像重建,适用于灾害监控中的大面积遥感数据处理[169,173]。未来研究可聚焦于优化编码算法的计算效率和实时性,尤其是针对当前大部分算法在编码时间上的不足,减少冗余计算。

6.4　高光谱数据的编码算法与应用

高光谱数据的编码任务中,码率 R、任务质量 Q 和可计算性 C 之间的关系较为复杂。其主要原因在于高光谱数据的每个像素通常包含数十到上百个波段的信息,导致其多波段特性和庞大数据量使得单一的编码方法难以在 R-Q-C 之间取得有效平衡。实验表明,在高光谱数据编码过程中,随着压缩率的提升,部分光谱和空间信息不可避免地丢失。例如,在 HCCNet[184]的研究中,当多波段比特率降至 0.228 时,PSNR 下降至 30.30 dB。尽管较低的比特率减少了传输带宽,但也带来了显著的重构精度损失。在极端压缩条件下,高光谱数据的多维特性使得保持高质量的重构变得更加困难。随着编码比率的增加,如何在确保带宽效率的同时最大限度地保持数据完整性,仍然是编码任务中的核心难题[182]。

高光谱编码算法在精准农业、遥感监测、灾害应急等应用中展现出显著优势。在精准农业中,DC-SN 算法[187]通过轻量化设计,优化了农业监测中的数据压缩和传输性能。基于压缩感知的 HyperLCA 算法[188]在低比特率下依然能保留丰富的光谱信息,适用于无人机平台上的作物监测任务,特别是在带宽有限的情况下进行实时作物健康分析。对于同样需要高效传输的应用,在灾害监测中,CENet 算法[25]的双注意力机制结合边缘引导策略,能够通过动态调整码率有效压缩地震、火山等高频区域数据,同时减少不重要区域的码率消耗,确保紧急情况下的高精度数据传输。在城市监测中,通过边缘引导的高效压缩策略,能够在带宽受限的条件下实时捕获和传输建筑物和基础设施的高光谱信息,有效支持城市规划和灾害预警中的高效数据处理[189]。此外,在环境遥感中,通过跨通道对比学习[179]能够在低比特率下仍能保持较高的重构质量,使其适用于广域遥感数据的长期监测。

7　未来的研究方向

机器视觉编码技术凭借其新颖的结构、较低的编码复杂度和良好的鲁棒性,正在逐渐成为一种重要的新型编码方法,吸引了国内外学术界的广泛关注。尽管已经取得了一些重要的研究成果,机器视觉编码技术仍然面临许多技术挑战,需要进一步深入研究和探索。未来的研究可以从以下几个方向展开:

(1)多传感器数据联合编码:面向未来的高端应用需求,如自动驾驶、医疗影像分析和元宇宙等,都

需要综合分析来自多种传感器的数据，如可见光、非可见光、3D 点云和多频谱信号。如何根据这些不同类型的信号特点生成不同的码流进行联合编码和运动估计，将是提高视觉任务完成质量的关键。

(2)特征压缩编码：随着大模型(如 DeepSeek[211]、Sora 文生视频[212])的兴起，不同时序模块产生了大量的视觉特征用于机器训练和推理，如何高效编码和压缩大量视觉特征是重大挑战。未来应致力于开发更高效的特征压缩编码方法，以减少数据传输和存储的成本，同时保持关键视觉信息。此外，大模型的图生文和文生图技术为特征压缩编码提供了新的思路，未来可通过生成性模型提升语义一致性和编码效率。

(3)语义表示方法：现有的机器视觉编码方法主要依赖特征提取和编码过程，但对于复杂的视觉场景和任务，传统的特征表示方法可能存在局限性。未来研究应改进语义表示方法，更好地捕捉视觉信息中的语义和上下文关系。此外，可以考虑结合大模型的语义理解能力，提升编码对复杂场景中隐含信息的把握和表示。

(4)人机混合编码：结合机器的高效处理能力和人类的感知与决策能力，实现更准确和智能的视觉编码和分析。这可能涉及开发交互式编码界面、设计有效的人机协作策略，以及解决人机交互中的挑战，如解释性和透明度问题。

(5)可伸缩性架构：在大规模应用场景中，机器视觉编码面临着可伸缩性的挑战。未来研究可以致力于提高机器视觉编码的可伸缩性，以应对大规模数据和复杂任务的需求。涉及设计分布式编码系统、开发高效的并行计算方法，以及优化编码算法以适应不同规模的计算资源和数据集。

(6)机器任务驱动的模型编码：随着人工智能技术的不断演进，传统视频编码方法已难以满足智能应用对数据处理效率与任务精准度的双重要求。机器任务驱动的模型编码通过将编码流程与后续视觉任务(如目标检测、图像分割等)紧密结合，采用任务驱动的设计策略，实现对关键视觉信息的高效提取与压缩。借助深度学习算法的持续迭代优化，成为未来智能视频处理的重要发展方向。

(7)研发专用视觉处理器：视觉处理器 VPU(Vision Processing Unit)是一种以机器视觉为核心、深度融合图像信号处理、数据分析、视频编码与 AI 计算的专用处理器，能够对可见光及非可见光电磁波(如红外、激光点云等)、物理传感信号以及合成数据等多模态信息进行智能化处理，并同步生成编码数据与决策结果。其核心优势在于通过硬件级加速实现高精度视觉分析、低功耗实时处理以及类脑化的“感知－认知”协同能力。当前主流 VPU 仍聚焦于人眼可感知的可见光视频处理(如多路高清编码等)，尚未充分释放机器视觉超越人眼的潜力。未来 VPU 设计将突破传统光谱限制，深度融合红外热成像、雷达、LiDAR 点云等更多机器视觉数据，构建高能效异构计算架构，集成动态分辨率映射、多模态特征融合引擎、多类视觉数据并行计算、存算一体单元及自主学习模块，支持高速信号处理(如微秒级事件驱动计算)、场景语义重建与实时推理决策，成为以视觉为中心的脑眼融合专用处理器。

机器视觉编码技术作为视频编码领域的前沿研究方向，通过创新性技术架构能够有效解决传统视频编码中存在的鲁棒性不足与高编码复杂度等核心问题。本文系统阐释了该技术体系的数据来源、理论基础及关键技术组成，重点剖析了主流编码框架的特征优势，通过技术演进脉络的梳理揭示其发展规律。在此基础上，本文进一步提出未来研究的重点方向与发展路径，旨在推动该技术在理论体系与工程实践层面的纵深发展及跨领域应用拓展，为相关领域的学者提供具有参考价值的研究视角。

参 考 文 献

[1] Yang Wen-Han, Huang Hao-Feng, Hu Yue-Yu, et al. Video coding for machine: compact visual representation compression for intelligent collaborative analytics. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(7): 1-14

[2] Zhao Tie-Song, Huang Yu-Hang, Feng Wei-Ze, et al. Efficient VVC intra prediction based on deep feature fusion and probability estimation. IEEE Transactions on Multimedia, 2022, 25: 6411-6421

[3] Ma Le, Wu Xiao-Yue, Li Zhi-Wei. High-precision medicine bottles vision online inspection system and classification based on multifeatures and ensemble learning via independence test. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-12

[4] Muzahid Abu Jafar Md, Kamarulzaman Syafiq Fauzi, Rahman Md Arafatur, et al. Multiple vehicle cooperation and collision avoidance in automated vehicles: survey and an ai-enabled conceptual framework. Nature Scientific Reports, 2023, 13(603): 1-27

[5] Awasthi Shubham, Jain Kamal, Sahoo Sashikanta, et al. Analyzing joshimath's sinking: causes, consequences, and future prospects with remote sensing techniques. Nature Scientific

Reports, 2024, 14(10876): 1-25

[6] Wang Jun-Ke, Chen Dong-Dong, Luo Chong, et al. Chatvideo: a tracklet-centric multimodal and versatile video understanding system. arXiv preprint arXiv: 2304.14407, 2023: 1-10

[7] Yoon Yong-Uk, Han Gyu-Woong, Lee Jooyoung, et al. An advanced multi-scale feature compression using selective learning strategy for video coding for machines//Proceedings of the IEEE International Conference on Visual Communications and Image Processing. Jeju, Republic of Korea, 2023: 1-5

[8] MPEG. Compact descriptors for visual search, https://mpeg.chiariglione.org/standards/mpeg-7/compact-descriptors-visual-search, 2015

[9] MPEG. Compact descriptors for video analysis, https://mpeg.chiariglione.org/standards/mpeg-7/compact-descriptors-video-analysis, 2019

[10] Duan Ling-Yu, Liu Jia-Ying, Yang Wen-Han, et al. Video coding for machines: a paradigm of collaborative compression and intelligent analytics. IEEE Transactions on Image Processing, 2020, 29: 8680-8695

[11] Wang Yu-Qin, Xu Jing-Ce, Ji Wen. A feature-based video transmission framework for visual IoT in fog computing systems//Proceedings of the Symposium on Architectures for Networking and Communications Systems. Cambridge, UK, 2019: 1-8

[12] Wang Yu-Qin, Xu Jing-Ce, Ji Wen. An optimal coverage model for the deployment of iot devices in feature-based video transmission systems//Proceedings of the IEEE Visual Communications and Image Processing. Sydney, Australia, 2019: 1-4

[13] Li Ming-Xuan, Ji Wen. Lightweight multiattention recursive residual cnn-based in-loop filter driven by neuron diversity. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(11): 6996-7008

[14] Liu Jian-Ran, Zhang Chang, Ji Wen. End-to-end image compression through machine semantics//Proceedings of the International Forum of Digital Multimedia Communication. Singapore, 2023: 1-12

[15] Li Ming-Xuan, Ji Wen. Screen content-aware video coding through non-local model embedded with intra inter in-loop filtering. IEEE Transactions on Circuits and Systems for Video Technology, 2025, 35(2): 1870-1883

[16] Ji Wen, Liang Bing, Wang Yu-Qin, et al. Crowd v-ioe: visual internet of everything architecture in ai-driven fog computing. IEEE Wireless Communications, 2020, 27(2): 51-57

[17] Park Jeongsoo, Johnson Justin. Rgb no more: minimally decoded jpeg vision transformers//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 22334-22346

[18] Wang Shu-Run, Wang Zhao, Wang Shi-Qi, et al. Deep image compression towards machine vision: a unified optimization framework. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 33(6): 2979-2989

[19] Ma Si-Wei, Zhang Li, Wang Shi-Qi, et al. Evolution of avs video coding standards: twenty years of innovation and development. Science China Information Sciences, 2022, 65(9): 1-24

[20] Hertz Heinrich. On very rapid electric oscillations. London, UK: Routledge, 2019: 193-222

[21] Fossum Eric R. The invention of CMOS image sensors: a camera in every pocket//Proceedings of the Pan Pacific Microelectronics Symposium. HI, USA, 2020: 1-6

[22] Gao Wen, Xu Xiao-Zhong, Qin Matthew, et al. An open dataset for video coding for machines standardization//Proceedings of the IEEE International Conference on Image Processing. Bordeaux, France, 2022: 4008-4012

[23] Wang Wei, Jin Zhi. Capformer: compression-aware pre trained transformer for low-light image enhancement//Proceedings of the International Conference on Multimedia and Expo. Niagara Falls, Canada, 2024: 1-6

[24] Systems FLIR. Free flir thermal dataset for algorithm training, https://www.flir.com/oem/adas/adas-dataset-form/, 2022

[25] Guo Yuan-Yuan, Tao Yu-Long, Chong Yan-Wen, et al. Edge-guided hyperspectral image compression with interactive dual attention. IEEE Transactions on Geoscience and Remote Sensing, 2022, 61: 1-17

[26] Behley Jens, Garbade Martin, Milioto Andres, et al. A dataset for semantic segmentation of point cloud sequences. arXiv preprint arXiv: 1904.01416, 2019, 2: 1-12

[27] Vitale Santiago, Orlando Jos′e Ignacio, Iarussi Emmanuel, et al. Improving realism in patient-specific abdominal ultrasound simulation using cyclegans. International Journal of Computer Assisted Radiology and Surgery, 2020, 15(2): 183-192

[28] Laboratories Sandia National. Sandia SAR data, https://www.sandia.gov/radar/complex-data/, 2020

[29] Liu Xing, Zhu Ming-Yu, Zheng Si-Ming, et al. Video snapshot compressive imaging using adaptive progressive coding for high-quality reconstruction under different illumination circumstances. Optics Letters, 2023, 49(1): 85-88

[30] Cai Shilv, Chen Liqun, Zhong Sheng, et al. Make lossy compression meaningful for low-light images//Proceedings of the AAAI Conference on Artificial Intelligence. Singapore, 2024: 8236-8245

[31] Bhowmik Neelanjan, Barker Jack W, Gaus Yona Falinie A, et al. Lost in compression: the impact of lossy image compression on variable size object detection within infrared imagery//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 369-378

[32] Wang Ruo-Fan, Mao Qi, Wang Shi-Qi, et al. Disentangled visual representations for extreme human body video compression//Proceedings of the IEEE International Conference

on Multimedia and Expo. Taipei, China, 2022: 1-6

[33] Gao Wen, Liu Shan, Xu Xiao-Zhong, et al. Recent standard development activities on video coding for machines. arXiv preprint arXiv:2105.12653, 2021: 1-13

[34] Zhang Qi, Wang Shan-She, Zhang Xin-Feng, et al. Just recognizable distortion for machine vision oriented image and video coding. International Journal of Computer Vision, 2021, 129(10): 2889-2906

[35] Veerabadran Vijay, Pourreza Reza, Habibian Amirhossein, et al. Adversarial distortion for learned video compression// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Seattle, USA, 2020: 640-644

[36] Le Nam, Zhang Honglei, Cricri Francesco, et al. Image coding for machines: an end-to-end learned approach//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto, Canada, 2021: 1590-1594

[37] Welstead Stephen T. Fractal and wavelet image compression techniques//Proceedings of the Storage and Retrieval for Image and Video Databases. San Jose, USA, 1999: 155-156

[38] Wang Zhou, Bovik Alan C, Sheikh Hamid R, et al. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 2004, 13(4): 600-612

[39] Zhao Long, Gundavarapu Nitesh B, Yuan Liang-Zhe, et al. Videoprism: a foundational visual encoder for video understanding. arXiv preprint arXiv:2402.13217, 2024: 1-26

[40] Jang Ho-Jin, Tong Frank. Improved modeling of human vision by incorporating robustness to blur in convolutional neural networks. Nature Communications, 2024, 15(1989): 1-14

[41] Bastug Ejder, Bennis Mehdi, Medard Muriel, et al. Toward interconnected virtual reality: opportunities, challenges, and enablers. IEEE Communications Magazine, 2017, 55(6): 110-117

[42] Zeng Huan-Qiang, Kong Qing-Wei, Chen Jing, et al. Immersive video coding technology review. Journal of Electronics and Information, 2024, 46(2): 602-614 (in Chinese)
(曾焕强, 孔庆玮, 陈婧, 等. 沉浸式视频编码技术综述. 电子与信息学报, 2024, 46(2): 602-614)

[43] MPEG-AI. Part 3: Optimization of encoders and receiving systems for machine analysis of coded video content, https://www.mpeg.org/standards/MPEG-AI/, 2024

[44] Bross Benjamin, Wang Ye-Kui, Ye Yan, et al. Overview of the versatile video coding (VVC) standard and its applications. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(10): 3736-3764

[45] Heinrich-Hertz-Institute Fraunhofer. Enhanced compression model, https://wiki.x266.mov/docs/video/ECM, 2024

[46] Zhang Jia-Qi, Jia Chuan-Min, Lei Meng, et al. Recent development of AVS video coding standard: AVS3//Proceedings of the Picture Coding Symposium. Ningbo, China, 2019: 1-5

[47] Graziosi Danillo, Nakagami Ohji, Kuma Satoru, et al. An overview of ongoing point cloud compression standardization activities: video-based (V-pcc) and geometry-based (G-pcc). APSIPA Transactions on Signal and Information Processing, 2020, 9(13): 1-17

[48] Wiegand Thomas, Sullivan Gary J, Bjontegaard Gisle, et al. Overview of the H.264/AVC video coding standard. IEEE Transactions on Circuits and Systems for Video Technology, 2003, 13(7): 560-576

[49] Qi Shu-Ren, Zhang Yu-Shu, Wang Chao, et al. A survey of orthogonal moments for image representation: theory, implementation, and evaluation. ACM Computing Surveys, 2021, 55(1): 1-35

[50] Chang Jian-Hui, Zhao Zheng-Hui, Yang Ling-Bo, et al. Thousand to one: semantic prior modeling for conceptual coding//Proceedings of the IEEE International Conference on Multimedia and Expo. Shenzhen, China, 2021: 1-6

[51] Jing Xin-Yi, Feng Qiao, Lai Yu-Kun, et al. State: learning structure and texture representations for novel view synthesis. Computational Visual Media, 2023, 9(4): 767-786

[52] Wang Su-Hong, Jia Chuan-Min, Zhang Xin-Feng, et al. A pixel-level segmentation-synthesis framework for dynamic texture video compression. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(10): 7077-7091

[53] Chang Jian-Hui, Zhao Zheng-Hui, Jia Chuan-Min, et al. Conceptual compression via deep structure and texture synthesis. IEEE Transactions on Image Processing, 2022, 31: 2809-2823

[54] Yang Guo-Ye, Nakayama George Kiyohiro, Xiao Zi-Kai, et al. Semantic-aware transformation-invariant RoI align//Proceedings of the AAAI Conference on Artificial Intelligence. Online, 2024, 38: 6486-6493

[55] He Ling-Feng, Xu Meng-Ze, Ma Jie. Weakly-supervised roi extraction method based on contrastive learning for remote sensing images//Proceedings of the IEEE International Geoscience and Remote Sensing Symposium. Pasadena, USA, 2023: 6378-6381

[56] Im Sio-Kei, Chan Ka-Hou. Cabac-based RoI encryption with mask R-CNN for VVC codec//Proceedings of the International Conference on Ubiquitous Information Management and Communication. Kuala Lumpur, Malaysia, 2024: 1-6

[57] Strinati Emilio Calvanese, Barbarossa Sergio. 6G networks: beyond Shannon towards semantic and goal oriented communications. Computer Networks, 2021, 190: 1-52

[58] Wang Ting-Chun, Mallya Arun, Liu Ming-Yu. One shot free-view neural talking-head synthesis for video conferencing//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 10039-10049

[59] Jin Xin, Feng Ruo-Yu, Sun Si-Meng, et al. Semantical video

coding: instill static-dynamic clues into structured bitstream for AI tasks. Journal of Visual Communication and Image Representation, 2023, 93: 1-21

[60] Guo Zong-Yu, Zhang Zhi-Zheng, Feng Run-Sen, et al. Causal contextual prediction for learned image compression. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(4): 2329-2341

[61] Liu Qian-Kun, Liu Bin, Wu Yue, et al. Real-time online multi-object tracking in compressed domain. arXiv preprint arXiv:2204.02081, 2022: 1-10

[62] Liu Shi-Zhan, Lin Wei-Yao, Chen Yi-Hang, et al. A unified framework for jointly compressing visual and semantic data. ACM Transactions on Multimedia Computing, Communications and Applications, 2024, 20(7): 1-24

[63] Tan Zhen-Tao, Liu Bin, Chu Qi, et al. Real time video object segmentation in compressed domain. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(1): 175-188

[64] Zhang Yu-Feng, Lin Wei-Yao, Dai Wen-Rui, et al. Scene graph lossless compression with adaptive prediction for objects and relations. arXiv preprint arXiv:2304.13359, 2023: 1-18

[65] Zhang Ping-Ping, Wang Shi-Qi, Wang Meng, et al. Rethinking semantic image compression: scalable representation with cross-modality transfer. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(8): 4441-4445

[66] Chen Ting, Kornblith Simon, Norouzi Mohammad, et al. A simple framework for contrastive learning of visual representations//Proceedings of the International Conference on Machine Learning. Online, 2020: 1597-1607

[67] Tu Han-Yue, Li Li, Zhou Wen-Gang, et al. Semantic scalable image compression with cross-layer priors//Proceedings of the ACM International Conference on Multimedia. Online, 2021: 4044-4052

[68] Zhang Ping, Xu Wen-Jun, Gao Hui, et al. Toward wisdom evolutionary and primitive-concise 6G: a new paradigm of semantic communication networks. Engineering, 2022, 8: 60-73

[69] Li Ji-Guo, Jia Chuan-Min, Zhang Xin-Feng, et al. Cross modal compression: towards human-comprehensible semantic compression//Proceedings of the ACM International Conference on Multimedia. Online, 2021: 4230-4238

[70] Bai Yuan-Chao, Yang Xu, Liu Xian-Ming, et al. Towards end-to-end image compression and analysis with transformers//Proceedings of the AAAI Conference on Artificial Intelligence. Online, 2022, 36: 104-112

[71] Xie Hui-Qiang, Qin Zhi-Jin, Tao Xiao-Ming, et al. Task oriented multi-user semantic communications. IEEE Journal on Selected Areas in Communications, 2022, 40(9): 2584-2597

[72] Lee Kuang-Huei, Arnab Anurag, Guadarrama Sergio, et al. Compressive visual representations. Advances in Neural Information Processing Systems, 2021, 34: 19538-19552

[73] Liu Ze, Lin Yu-Tong, Cao Yue, et al. Swin transformer: hierarchical vision transformer using shifted windows//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 10012-10022

[74] Tan Ming-Xing, Le Quoc. Efficientnetv2: smaller models and faster training//Proceedings of the International Conference on Machine Learning. Online, 2021: 10096-10106

[75] Yu Wei-Hao, Zhou Pan, Yan Shui-Cheng, et al. Inceptionnext: when inception meets convnext//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 5672-5683

[76] Li An-Di, Ying Yu-Han, Gao Tian, et al. Mf net: multiscale feature extraction-integration network for unsupervised deformable registration. Frontiers in Neuroscience, 2024, 18: 1-10

[77] Wood Daniel. Task oriented video coding: a survey. arXiv preprint arXiv:2208.07313, 2022: 1-9

[78] Brummer Benoit, De Vleeschouwer Christophe. Adapting jpeg xs gains and priorities to tasks and contents//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, USA, 2020: 164-165

[79] Huang Zhi-Meng, Jia Chuan-Min, Wang Shan-She, et al. Visual analysis motivated rate-distortion model for image coding//Proceedings of the IEEE International Conference on Multimedia and Expo. Shenzhen, China, 2021: 1-6

[80] Li Xin, Shi Jun, Chen Zhi-Bo. Task-driven semantic coding via reinforcement learning. IEEE Transactions on Image Processing, 2021, 30: 6307-6320

[81] Zhao Yu, Luo Deng-Yan, Wang Fu-Chun, et al. End-to end compression for surveillance video with unsupervised foreground-background separation. IEEE Transactions on Broadcasting. 2023, 69(4): 966-978

[82] Chamain Lahiru D, Racap′e Fabien, B′egaint Jean, et al. End-to-end optimized image compression for machines, a study//Proceedings of the Data Compression Conference. Snowbird, USA, 2021: 163-172

[83] Wang Chien-Yao, Yeh I-Hau, Liao Hong-Yuan Mark. Yolo v9: learning what you want to learn using programmable gradient information//Proceedings of the European conference on computer vision. Milan, Italy, 2024: 1-18

[84] Wang Ao, Chen Hui, Liu Li-Hao, et al. Yolo v10: real time end-to-end object detection. Advances in Neural Information Processing Systems, 2024, 37: 107984-108011

[85] Cheng Tian-Heng, Song Lin, Ge Yi-Xiao, et al. Yolo world: real-time open-vocabulary object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 16901-16911

[86] Alvar Saeed Ranjbar, Baji′c Ivan V. Pareto-optimal bit allocation for collaborative intelligence. IEEE Transactions on Image Processing, 2021, 30: 3348-3361

[87] Hu Yue-Yu, Yang Wen-Han, Huang Hao-Feng, et al. Re-

visit visual representation in analytics taxonomy: a compression perspective. arXiv preprint arXiv: 2106.08512, 2021: 1-11

[88] Lee Yegi, Kim Shin, Yoon Kyoungro, et al. Machine attention-based video coding for machines//Proceedings of the IEEE International Conference on Image Processing. Kuala Lumpur, Malaysia, 2023: 2700-2704

[89] Ladune Th′eo, Philippe Pierrick, Hamidouche Wassim, et al. Conditional coding and variable bitrate for practical learned video coding. arXiv preprint arXiv:2104.09103, 2021: 1-5

[90] Fischer Kristian, Brand Fabian, Herglotz Christian, et al. Video coding for machines with feature-based rate distortion optimization//Proceedings of the IEEE International Workshop on Multimedia Signal Processing. Tampere, Finland, 2020: 1-6

[91] Gou Ao-Rui, Sun He-Ming, Zeng Xiao-Yang, et al. Fast vvc intra encoding for video coding for machines//Proceedings of the IEEE International Symposium on Circuits and Systems. Monterey, USA, 2023: 1-5

[92] Yılmaz M Akın, Tekalp A Murat. End-to-end rate distortion optimized learned hierarchical bi-directional video compression. IEEE Transactions on Image Processing, 2021, 31: 974-983

[93] Zhang Yi-Wei, Lu Guo, Chen Yu-Nuo, et al. Neural rate control for learned video compression//International Conference on Learning Representations. Singapore, 2024: 1-15

[94] Ge Xing-Tong, Luo Ji-Xiang, Zhang Xin-Jie, et al. Task-aware encoder control for deep video compression//IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 26036-26045

[95] Cheng Ho Kei, Schwing Alexander G. Xmem: long term video object segmentation with an atkinson-shiffrin memory model//European Conference on Computer Vision. Online, 2022: 640-658

[96] Pernu˘s Martin, ˘Struc Vitomir, Dobri˘sek Simon. Maskface-GAN: high-resolution face editing with masked GAN latent code optimization. IEEE Transactions on Image Processing, 2023, 32: 5893-5908

[97] Feng Ruo-Yu, Jin Xin, Guo Zong-Yu, et al. Image coding for machines with omnipotent feature learning//Proceedings of the European Conference on Computer Vision. Online, 2022: 510-528

[98] Alvar Saeed Ranjbar, Baji′c Ivan V. Multi-task learning with compressible features for collaborative intelligence//Proceedings of the IEEE International Conference on Image Processing. Taipei, China, 2019: 1705-1709

[99] Qin Zhi-Jin, Gao Fei-Fei, Lin Bo, et al. A generalized semantic communication system: from sources to channels. IEEE Wireless Communications, 2023, 30(3): 18-26

[100] Dai Jin-Cheng, Zhang Ping, Niu Kai, et al. Communication beyond transmitting bits: semantics guided source and channel coding. IEEE Wireless Communications, 2022, 30(4): 170-177

[101] Luo Xue-Wen, Chen Hsiao-Hwa, Guo Qing. Semantic communications: overview, open issues, and future research directions. IEEE Wireless Communications, 2022, 29(1): 210-219

[102] Shi Guang-Ming, Xiao Yong, Li Ying-Yu, et al. From semantic communication to semantic-aware networking: model, architecture, and open problems. IEEE Communications Magazine, 2021, 59(8): 44-50

[103] Wang Shu-Run, Yang Wen-Han, Wang Shi-Qi. End-to end facial deep learning feature compression with teacher-student enhancement. arXiv preprint arXiv:2002.03627, 2020: 1-5

[104] Zhu Chen, Huang Yan, Xie Rong, et al. HEVC VMAF oriented perceptual rate distortion optimization using CNN//Proceedings of the Picture Coding Symposium. Bristol, UK, 2021: 1-5

[105] Agustsson Eirikur, Minnen David, Johnston Nick, et al. Scale-space flow for end-to-end optimized video compression//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 8503-8512

[106] Habibian Amirhossein, Rozendaal Ties van, Tomczak Jakub M, et al. Video compression with rate-distortion autoencoders//Proceedings of the IEEE/CVF International Conference on Computer Vision. Online, 2019: 7033-7042

[107] Liu Dong, Li Yue, Lin Jian-Ping, et al. Deep learning based video coding: a review and a case study. ACM Computing Surveys, 2020, 53(1): 1-35

[108] Qiu Han, Zheng Qin-Kai, Memmi Gerard, et al. Deep residual learning-based enhanced JPEG compression in the internet of things. IEEE Transactions on Industrial Informatics, 2020, 17(3): 2124-2133

[109] Li Jia-Hao, Li Bin, Lu Yan. Deep contextual video compression. Advances in Neural Information Processing Systems, 2021, 34: 18114-18125

[110] Sheng Xi-Hua, Li Jia-Hao, Li Bin, et al. Temporal context mining for learned video compression. IEEE Transactions on Multimedia, 2023, 25: 7311-7322

[111] Wu Li-Rong, Huang Ke-Jie, Shen Hai-Bin, et al. Foreground-background parallel compression with residual encoding for surveillance video. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(7): 2711-2724

[112] Liu Yu-Tong, Kong Ling-He, Chen Gui-Hai, et al. Lightweight AI and IoT collaboration for surveillance video preprocessing. Journal of Systems Architecture, 2021, 114: 1-13

[113] Chen Zhuo, Fan Kui, Wang Shi-Qi, et al. Toward intelligent sensing: intermediate deep feature compression. IEEE Transactions on Image Processing, 2020, 29: 2230-2243

[114] Suzuki Satoshi, Takeda Shoichiro, Takagi Motohiro, et al. Deep feature compression using spatio-temporal arrange-

ment toward collaborative intelligent world. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(6): 3934-3946

[115] Lee Jin-Young, Choi Yongho, Van Le The, et al. Efficient feature coding based on performance analysis of versatile video coding (VVC) in video coding for machines (VCM). Multimedia Tools and Applications, 2023, 82(27): 42803-42816

[116] Duarte Adson, Zatt Bruno, Correa Guilherme, et al. Fast intra mode decision using machine learning for the versatile video coding standard//Proceedings of the IEEE International Symposium on Circuits and Systems. Monterey, USA, 2023: 1-5

[117] Chang Shih-Fu, Sikora Thomas, Purl Atul. Overview of the MPEG-7 standard. IEEE Transactions on Circuits and Systems for Video Technology, 2001, 11(6): 688-695

[118] Ahonen Jukka I., Le Nam, Zhang Hong-Lei, et al. NN-VVC: versatile video coding boosted by self-supervisedly learned image coding for machines//Proceedings of the IEEE International Symposium on Multimedia. Laguna Hills, USA, 2023: 10-19

[119] Liu Hao-Jie, Lu Ming, Chen Zhi-Qi, et al. End-to-end neural video coding using a compound spatiotemporal representation. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(8): 5650-5662

[120] Meng Xue-Wei, Jia Chuan-Min, Zhang Xin-Feng, et al. Spatio-temporal correlation guided geometric partitioning for versatile video coding. IEEE Transactions on Image Processing, 2021, 31: 30-42

[121] Lee So Yoon, Yang Yoonmo, Kim Dongsin, et al. Offset based in-loop filtering with a deep network in HEVC. IEEE Access, 2020, 8: 213958-213967

[122] Chadha Aaron, Andreopoulos Yiannis. Deep perceptual preprocessing for video coding//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 14852-14861

[123] Yang Run-Yu, Liu Dong, Ma Si-Wei, et al. Knowledge distillation from end-to-end image compression to VVC intra coding for perceptual quality enhancement//Proceedings of the IEEE International Conference on Image Processing. Anchorage, USA, 2021: 3438-3442

[124] Wang Da-Yong, Lu Xin, Sun Yu, et al. A probability based zero-block early termination algorithm for QSHVC. IEEE Transactions on Broadcasting, 2023, 69(2): 1-13

[125] Gao Wen, Ma Si-Wei, Duan Ling-Yu, et al. Digital retina: a way to make the city brain more efficient by visual coding. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(11): 4147-4161

[126] Liu Kang, Liu Dong, Li Li, et al. Semantics-to-signal scalable image compression with learned reversible representations. International Journal of Computer Vision, 2021, 129(9): 2605-2621

[127] Sheng Xi-Hua, Li Li, Liu Dong, et al. VNVC: a versatile neural video coding framework for efficient human-machine vision. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(7): 4579-4596

[128] Codevilla Felipe, Simard Jean Gabriel, Goroshin Ross, et al. Learned image compression for machine perception. arXiv preprint arXiv:2111.02249, 2021: 1-13

[129] Lu Guo, Ge Xing-Tong, Zhong Tian-Xiong, et al. Preprocessing enhanced image compression for machine vision. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(12): 1-12

[130] Lu Guo, Zhong Tian-Xiong, Geng Jing, et al. Learning based multi-modality image and video compression//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 6083-6092

[131] Li Ang, Wei Xin, Wu Dan, et al. Cross-modal semantic communications. IEEE Wireless Communications, 2022, 29(6): 144-151

[132] Chang Jian-Hui, Mao Qi, Zhao Zheng-Hui, et al. Layered conceptual image compression via deep semantic synthesis//Proceedings of the IEEE International Conference on Image Processing. Taipei, China, 2019: 694-698

[133] Singh Saurabh, Abu-El-Haija Sami, Johnston Nick, et al. End-to-end learning of compressible features//Proceedings of the IEEE International Conference on Image Processing. Abu Dhabi, United Arab Emirates, 2020: 3349-3353

[134] Lan Qiao, Wen Ding-Zhu, Zhang Ze-Zhong, et al. What is semantic communication? a view on conveying meaning in the era of machine intelligence. Journal of Communications and Information Networks, 2021, 6(4): 336-374

[135] Hu Yue-Yu, Yang Shuai, Yang Wen-Han, et al. Towards coding for human and machine vision: a scalable image coding approach//Proceedings of the IEEE International Conference on Multimedia and Expo. London, UK, 2020: 1-6

[136] Wang Jin-Peng, Gao Yu-Ting, Li Ke, et al. Enhancing unsupervised video representation learning by decoupling the scene and the motion//Proceedings of the AAAI Conference on Artificial Intelligence. Online, 2021: 10129-10137

[137] Xie Zhen-Da, Lin Yu-Tong, Zhang Zheng, et al. Propagate yourself: exploring pixel-level consistency for unsupervised visual representation learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 16684-16693

[138] Dilip Rohit, Liu Yu-Jie, Smith Adam, et al. Data compression for quantum machine learning. Physical Review Research, 2022, 4(4): 1-8

[139] Lin Jie, Duan Ling-Yu, Wang Shi-Qi, et al. HNIP: compact deep invariant representations for video matching, localization, and retrieval. IEEE Transactions on Multimedia, 2017, 19(9): 1968-1983

[140] Ding Lin, Tian Yong-Hong, Fan Hong-Fei, et al. Joint

coding of local and global deep features in videos for visual search. IEEE Transactions on Image Processing, 2020, 29: 3734-3749

[141] Lamghari Soufiane, Bilodeau Guillaume-Alexandre, Saunier Nicolas. Actar: actor-driven pose embeddings for video action recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 399-408

[142] Kang Xu, Song Bin, Guo Jie, et al. Task-oriented image transmission for scene classification in unmanned aerial systems. IEEE Transactions on Communications, 2022, 70(8): 5181-5192

[143] Krizhevsky Alex, Sutskever Ilya, Hinton Geoffrey E. Imagenet classification with deep convolutional neural networks. Communications of the ACM, 2017, 60(6): 84-90

[144] Simonyan Karen, Zisserman Andrew. Very deep convolutional networks for large-scale image recognition. Online, 2014: 1-14

[145] Szegedy Christian, Liu Wei, Jia Yang-Qing, et al. Going deeper with convolutions//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1-9

[146] He Kai-Ming, Zhang Xiang-Yu, Ren Shao-Qing, et al. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[147] Ren Shao-Qing, He Kai-Ming, Girshick Ross, et al. Faster R-CNN: towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems, 2015, 28: 1-9

[148] Vandenhende Simon, Georgoulis Stamatios, Van Gansbeke Wouter, et al. Multi-task learning for dense prediction tasks: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(7): 3614-3633

[149] He Kai-Ming, Gkioxari Georgia, Dollár Piotr, et al. Mask R-CNN//Proceedings of the IEEE International Conference on Computer Vision. Online, 2017: 2961-2969

[150] Karras Tero, Laine Samuli, Aittala Miika, et al. Analyzing and improving the image quality of stylegan//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 8110-8119

[151] Lou Yi-Hang, Duan Ling-Yu, Luo Yong, et al. Towards efficient front-end visual sensing for digital retina: a model-centric paradigm. IEEE Transactions on Multimedia, 2020, 22(11): 3002-3013

[152] Pernuš Martin, Štruc Vitomir, Dobrišek Simon. Maskface-GAN: high resolution face editing with masked GAN latent code optimization. IEEE Transactions on Image Processing, 2023, 32: 5893-5908

[153] Chen Jia-Wei, Ho Chiu-Man. MM-ViT: multi modal video transformer for compressed video action recognition//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA, 2022: 1910-1921

[154] Ma Si-Wei, Gao Jun-Long, Wang Ruo-Fan, et al. Overview of intelligent video coding: from model-based to learning-based approaches. Visual Intelligence, 2023, 1(1): 1-19

[155] Que Zi-Zheng, Lu Guo, Xu Dong. Voxelcontext-net: an octree based framework for point cloud compression//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 6038-6047

[156] He Yun, Ren Xin-Lin, Tang Dan-Hang, et al. Density preserving deep point cloud compression//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 2333-2342

[157] Akhtar Anique, Li Zhu, Van der Auwera Geert. Inter frame compression for dynamic point cloud geometry coding. IEEE Transactions on Image Processing, 2024, 33: 584-594

[158] Song Rui, Fu Chun-Yang, Liu Shan, et al. Efficient hierarchical entropy model for learned point cloud compression//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 14368-14377

[159] Fu Chun-Yang, Li Ge, Song Rui, et al. Octattention: octree-based large-scale contexts model for point cloud compression//Proceedings of the Association for the Advancement of Artificial Intelligence. Online, 2022: 1-10

[160] Wei Lei, Wan Shuai, Wang Zhe-Cheng, et al. Near lossless compression of point cloud attribute using quantization parameter cascading and rate-distortion optimization. IEEE Transactions on Multimedia, 2024, 26: 3317-3330

[161] Garcia Diogo C., Fonseca Tiago A., Ferreira Renan U., et al. Geometry coding for dynamic voxelized point clouds using octrees and multiple contexts. IEEE Transactions on Image Processing, 2020, 29: 313-322

[162] Cao Chao, Preda Marius, Zakharchenko Vladyslav, et al. Compression of sparse and dense dynamic point clouds—methods and standards//Proceedings of the IEEE. Online, 2021, 109: 1537-1558

[163] Song Fei, Li Ge, Yang Xiao-Dong, et al. Block-adaptive point cloud attribute coding with region-aware optimized transform. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(8): 4294-4308

[164] Krivoku′ca Maja, Chou Philip A., Koroteev Maxim. A volumetric approach to point cloud compression—part Ⅱ: geometry compression. IEEE Transactions on Image Processing, 2020, 29: 2217-2229

[165] Wang Jian-Qiang, Ding Dan-Dan, Li Zhu, et al. Sparse tensor-based multiscale representation for point cloud geometry compression. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(7): 9055-9071

[166] Li Xin, Dai Wen-Rui, Li Shao-Hui, et al. 3-D point cloud attribute compression with Laplacian embedding graph dic-

tionary learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(2): 975-993

[167] Quach Maurice, Valenzise Giuseppe, Dufaux Frederic. Folding-based compression of point cloud attributes//Proceedings of the IEEE International Conference on Image Processing. Abu Dhabi, United Arab Emirates, 2020: 3309-3313

[168] Zhang Li-Li, Pan Tian-Peng, Huang Yu-Feng, et al. Sar image compression using discretized gaussian adaptive model and generalized subtractive normalization. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 1-5

[169] Foix-Colonier Nils, Amao-Oliva Joel, Sica Francescopaolo. Deep learning-based compression and despeckling of sar images//Proceedings of the European Conference on Synthetic Aperture Radar. Munich, Germany, 2024: 237-242

[170] Fu Chuan, Du Bo, Zhang Liang-Pei. Sar image compression based on multi-resblock and global context. IEEE Geoscience and Remote Sensing Letters, 2023, 20: 1-5

[171] Pilikos Georgios, Azcueta Mario, Camarero Roberto, et al. Raw sar data compression with deep learning//Proceedings of the IEEE International Geoscience and Remote Sensing Symposium. Athens, Greece, 2024: 2546-2549

[172] Xu Qi-Han, Xiang Yun-Fan, Di Zhi-Xiong, et al. Synthetic aperture radar image compression based on a variational autoencoder. IEEE Geoscience and Remote Sensing Letters, 2021, 19: 1-5

[173] Fan Chun-Xiao, Hu Zhou, Jia Lu, et al. A novel lossless compression encoding framework for sar remote sensing images. Signal, Image and Video Processing, 2021, 15(3): 441-448

[174] Romano Diego, Lapegna Marco, Mele Valeria, et al. Designing a GPU-parallel algorithm for raw SAR data compression: a focus on parallel performance estimation. Future Generation Computer Systems, 2020, 112: 695-708

[175] Scheiber Rolf, Martone Michele, Gollin Nicola. Chirp selection and data compression for spaceborne wide swath SAR in FSCAN-mode//European Conference on Synthetic Aperture Radar. Online, 2021: 1-6

[176] Hay Craig, Donnell Lucy, Crawshaw Charlotte, et al. Adaptive on-board signal compression for SAR using machine learning methods//Annual Small Satellite Conference. Online, 2023: 1-10

[177] Hu Xian-Yang, Ma Chang-Zheng, Lu Xing-Yu, et al. Compressive sensing SAR imaging algorithm for LFMCW systems. IEEE Transactions on Geoscience and Remote Sensing, 2021, 59(10): 8486-8500

[178] Pestel-Schiller Ulrike, Ostermann Joern. Interpreter-based evaluation of compressed SAR images using JPEG and HEVC intra coding: compression can improve usability//Proceedings of the European Conference on Synthetic Aperture Radar. Online, 2021: 1-6

[179] La Grassa Riccardo, Re Cristina, Cremonese Gabriele, et al. Hyperspectral data compression using fully convolutional autoencoder. Remote Sensing, 2022, 14(10): 1-14

[180] Kuester Jannick, Gross Wolfgang, Middelmann Wolfgang. 1D-convolutional autoencoder based hyperspectral data compression. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2021, 43: 15-21

[181] Dua Yaman, Singh Ravi Shankar, Parwani Kshitij, et al. Convolutional neural network based lossy compression of hyperspectral images. Signal Processing: Image Communication, 2021, 95: 1-10

[182] Afrin Afsana, Al Mamun Md. A comprehensive review of deep learning methods for hyperspectral image compression//Proceedings of the International Conference on Advancement in Electrical and Electronic Engineering, Gazipur, Bangladesh, 2024: 1-6

[183] Wildenstein Diego, George Alan D. Towards intelligent compression of hyperspectral imagery//Proceedings of the IEEE International Conference on Electronics, Computing and Communication Technologies. Online, 2021: 1-6

[184] Guo Yuan-Yuan, Chong Yan-Wen, Pan Shao-Ming. Hyperspectral image compression via cross-channel contrastive learning. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-18

[185] Dua Yaman, Kumar Vinod, Singh Ravi Shankar. Comprehensive review of hyperspectral image compression algorithms. Optical Engineering, 2020, 59(9): 1-10

[186] Altamimi Amal, Ben Youssef Belgacem. A systematic review of hardware-accelerated compression of remotely sensed hyperspectral images. Sensors, 2021, 22(1): 1-53

[187] Hsu Chih-Chung, Lin Chia-Hsiang, Kao Chi-Hung, et al. DCSN: deep compressed sensing network for efficient hyperspectral data transmission of miniaturized satellite. IEEE Transactions on Geoscience and Remote Sensing, 2020, 59(9): 7773-7789

[188] Melián Jose, Díaz María, Morales Alejandro, et al. A novel data reutilization strategy for real-time hyperspectral image compression. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 1-5

[189] Das Samiran. Hyperspectral image, video compression using sparse Tucker tensor decomposition. IET Image Processing, 2021, 15(4): 964-973

[190] Lu Guo, Ouyang Wan-Li, Xu Dong, et al. DVC: an end-to-end deep video compression framework//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Online, 2019: 11006-11015

[191] Chen J., Ye Y., Kim S. JVET-Q2002-v3: algorithm description for versatile video coding and test model 8 (VTM 8)//Journal of Vibration Engineering and Technologies. Macao, China, 2020: 7-17

[192] Fan Yi-Bo. Principles of Video Codec Chip Design. Beijing: Science Press, 2022

范益波. 视频编解码芯片设计原理. 北京：科学出版社，2022

[193] Mao Qi, Wang Chong-Yu, Wang Meng, et al. Scalable face image coding via StyleGAN prior: towards compression for human-machine collaborative vision. IEEE Transactions on Image Processing, 2023, 33: 408-422

[194] Sandryhaila Aliaksei, Moura José MF. Discrete signal processing on graphs: graph Fourier transform//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada, 2013: 6167-6170

[195] Gastpar Michael, Dragotti Pier Luigi, Vetterli Martin. The distributed Karhunen-Loève transform. IEEE Transactions on Information Theory, 2006, 52(12): 5177-5196

[196] Wang Jian-Qiang, Ma Zhan. Sparse tensor-based point cloud attribute compression//Proceedings of the IEEE International Conference on Multimedia Information Processing and Retrieval. USA, 2022: 59-64

[197] Zamir Amir R, Sax Alexander, Shen William, et al. Taskonomy: disentangling task transfer learning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake, USA, 2018: 3712-3722

[198] Lin Tsung-Yi, Maire Michael, Belongie Serge, et al. Microsoft COCO: common objects in context//Proceedings of the European Conference on Computer Vision. Online, 2014: 740-755

[199] Xu Xiao-Zhong, Liu Shan, Li Ze-Qiang. A video dataset for learning-based visual data compression and analysis//Proceedings of the Conference on Visual Communications and Image Processing. Munich, Germany, 2021: 1-4

[200] Cordts Marius, Omran Mohamed, Ramos Sebastian, et al. The cityscapes dataset for semantic urban scene understanding//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 3213-3223

[201] Lee Cheng-Han, Liu Zi-Wei, Wu Ling-Yun, et al. MaskGAN: towards diverse and interactive facial image manipulation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 5549-5558

[202] Maji Subhransu, Rahtu Esa, Kannala Juho, et al. Fine grained visual classification of aircraft. arXiv preprint arXiv: 1306.5151, 2013: 1-6

[203] Parker Paul A, Holan Scott H, Ravishanker Nalini. Nonlinear time series classification using bispectrum based deep convolutional neural networks. Applied Stochastic Models in Business and Industry, 2020, 36(5): 877-890

[204] Deng Jia, Dong Wei, Socher Richard, et al. Imagenet: a large-scale hierarchical image database//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 248-254

[205] Sun Shi-Jie, Akhtar Naveed, Song Huan-Sheng, et al. Deep affinity network for multiple object tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 43(1): 104-119

[206] Behley Jens, Garbade Martin, Milioto Andres, et al. Semantickitti: a dataset for semantic scene understanding of lidar sequences//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 9297-9307

[207] Tulvan C, Gabrielli A, Preda M. Datasets update on point cloud compression for cultural objects. ISO/IEC JTC1/SC29 WG11 MPEG input document m38678, 2016: 1-12

[208] Pitz Wolfgang, Miller David. The Terrasar-x satellite. IEEE Transactions on Geoscience and Remote Sensing, 2010, 48(2): 615-622

[209] Yokoya, Naoto, Akira Iwasaki. Airborne hyperspectral data over Chikusei. Online, 2016, 5(5):1-6

[210] Piccioni G, Drossart P, Suetta Eea, et al. VIRTIS: the visible and infrared thermal imaging spectrometer. ESA Special Publication, 2007, 1295: 1-27

[211] DeepSeek-AI. DeepSeek,https://www.deepseek.com/, 2025

[212] Brooks Tim, Peebles Bill, Holmes Connor, et al. Video generation models as world simulators, https://openai.com, 2024

TIAN Gang-Yi, Ph. D. candidate. Her research interests include machine vision coding, image signal processing (ISP), and low-light algorithms.

JI Wen, Professor. Her research interests include vision processors, multimedia systems, and AI for Industries. She focuses on vision processing units, AI processor and systems for industries, multimedia end-edge-cloud systems, vision coding and transmission, intelligent multimedia computing, low-carbon computing, media economics, and optimization theories and methods.

Background

This research falls within the domain of data coding, specifically focusing on the efficient coding and processing of machine vision data, referred to as machine vision coding. Significant progress has been made in the coding of visible light data, but the coding techniques for infrared, ultraviolet, and point cloud data still require further exploration. In-

ternationally, the development of these technologies is still in its early stages, with general standards being formulated and no mature applications available yet.

This paper aims to delve deeply into machine vision coding, with a focus on data sources, fundamental principles, key technologies, and coding structures. It analyzes the development trajectory of this technology and looks forward to future research directions. The importance of this paper lies in attracting more scholars, both domestic and international, to participate in this field, thereby fostering more in-depth and extensive theoretical and applied research. The research team has achieved several notable results in the area of machine vision coding. Significant progress has been made in feature coding and human-machine hybrid vision coding systems, and this paper proposes a typical machine vision coding framework.

The relevance of this research is underscored by the increasing demand for efficient machine vision technology across various fields, such as industrial manufacturing, aerospace, marine exploration, and intelligent transportation. As the demand for high-quality machine vision tasks grows, so does the necessity for advanced compression coding technologies that can efficiently handle multiple data types. This research aims to provide practical solutions widely applicable to industry, thereby promoting the large-scale development and application of machine vision coding technology.