数字视频编解码技术综述

侯理想 聂婷屹 吴汶轩 李沈逸飞

摘要：视频编解码技术是数字视觉通信的基石，其发展经历了从传统混合编码到深度学习驱动的智能编码的演进。本文系统回顾了 H.26x 系列标准的迭代路径，阐述了其在压缩效率、网络适应性与应用场景方面的持续进步。同时，重点分析了基于深度学习的视频编码方法，包括其从架构模仿到条件生成、概率建模乃至 Transformer 与隐式表示等前沿范式的突破。研究表明，神经编码在多项性能指标上已接近或超越最新传统标准，并展现出向感知优化、内容自适应方向发展的趋势，为下一代沉浸式媒体应用提供了新的技术可能。

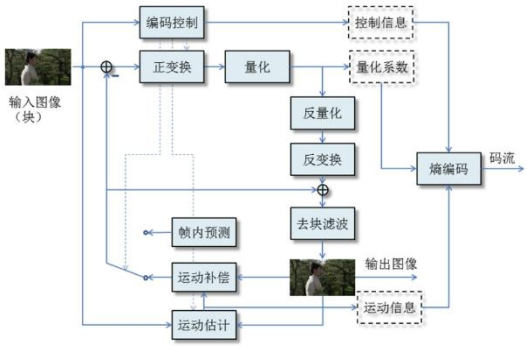
关键词：视频编码；混合编码框架；H.26x；深度学习；神经视频压缩；Transformer；率失真优化

1、引言

随着全球数字化进程的加速，视频数据已逐渐成为互联网流量的主体，占据了全球网络带宽的绝大部分份额。数字视频编解码技术作为多媒体通信系统的核心底层架构，其本质是为了解决海量视频数据在有限带宽和存储资源下的传输与保存难题。在未经压缩的原始状态下，数字视频信号携带的数据量极为惊人，以目前主流的超高清分辨率为例，短短几秒钟的无损画面便足以填满巨大的存储空间，这使得未经处理的视频流在现有的通信基础设施上几乎无法传输。因此，视频编解码技术通过一套严密的数学模型与信号处理算法，在编码端去除视频图像中的空间冗余、时间冗余、视觉冗余及编码冗余，将其压缩为紧凑的比特流；并在解码端通过逆向过程高保真地重建图像。这一技术不仅是流媒体服务、高清广播、远程会议及沉浸式媒体等应用存在的物理基础，更是降低数字化运营成本、提升用户视觉体验的关键所在。

传统的视频编解码技术主要基于经典的混合编码框架，该框架经过几十年的迭代已趋于成熟。其核心机制主要由预测、变换、量化和熵编码四个模块紧密耦合而成。具体而言，编码器首先通过帧内预测消除图像的

空间相关性，利用帧间预测和运动估计消除连续帧之间的时间相关性，从而得到残差信号；随后，通过离散余弦变换等方法将残差信号从像素域转换至频率域，分离高频与低频信息；接着，利用人类视觉系统对高频细节不敏感的特性，对变换系数进行量化处理，实现有损压缩；最后，通过算术编码等熵编码技术进一步去除统计冗余。这一系列复杂的操作旨在在保证感知画质的前提下，尽可能地降低码率。



在标准演进的历程中，H.264/AVC 标准无疑是过去二十年最成功的典范。它凭借适中的计算复杂度与卓越的兼容性，构筑了当前数字视频庞大的生态基石，至今仍广泛应用于各类终端与网络环境中。然而，随着 4K、8K 超高清视频以及 HDR 高动态范围内容的普及，H.264 的压缩效率已显疲态，促使了 H.265/HEVC 标准的诞生。H.265 通过引入更灵活的编码单元结构，将压缩效率提升了约 50%，成为了超高清时代的通用标准。与此同时，为了应对复杂的专利授权问题并适应互联网开放生态，Google 主导的 VP9 及由开放媒体联盟推出的 AV1 标准异军突起。AV1 以开源免版费为核心优势，虽然其编码复杂度较高，但凭借优秀的压缩性能和科技巨头的联合推动，正在迅速重塑流媒体分发的格局。此外，面向未来的 H.266/VVC 以及中国自主研发的 AVS3 标准，更是进一步将压缩极限推向了新的高度，为 8K 广播及 VR/AR 应用做好了技术储备。

当前，视频编解码领域正迎来一次具有颠覆性意义的技术范式转移，即基于深度学习的视频编解码技术的兴起。传统的混合编码框架依赖于人工设计的数学模型，经过数十年的优化，其性能提升已逐渐逼近香农极限，边际效应递减明显。而深度学习技术引入了数据驱动的理念，利用卷积神经网络（CNN）、循环神经网络（RNN）及变分自编码器（VAE）等架构，展现出了超越传统算法的潜力。基于深度学习的编解码主要分为两个研究方向：一是端到端的神经视频编码，即完全摒弃传统模块，通过神经网络直接学习图像到比特流的非线性映射，实现全局最优的压缩策略；二是混合架构增强，即保留传统框架，但在帧内预测、运动补偿或环路滤波等特定模块中引入 AI 模型，以提升局部处理的精度。例如，利用神经网络进行超分辨率重建或伪影去除，能显著提升主观视觉质量。

尽管基于深度学习的编解码技术在压缩效率上展现出广阔前景，但其目前仍面临计算复杂度极高、跨硬件平台通用性差以及缺乏统一标准化等挑战。然而，随着神经网络处理

单元（NPU）等专用硬件算力的提升以及MPEG 等标准化组织对神经网络视频编码（NNVC）探索的深入，AI 与编解码技术的深度融合已成为不可逆转的趋势。未来，视频编解码技术将不再局限于传统的信号处理范畴，而是向着智能化、感知化和内容自适应的方向演进，为元宇宙、全息通信等下一代沉浸式应用提供更加强大的底层动力。

本文将简单介绍传统混合编解码框架特别是 H.26x 系列标准的历史脉络，以及一些前沿的基于深度学习的视频编解码方式。

2、传统混合编码框架[1]-[21]

2.1 传统混合编码框架

传统混合编码框架是数字视频压缩领域的基础性技术架构，其核心设计围绕“预测 - 变换量化 - 熵编码” 的串联流程展开，通过充分挖掘视频数据的冗余特性实现高效压缩，同时搭配环路滤波等优化模块，在保证画质的前提下最大化降低数据量。这一框架的核心逻辑是通过分步处理逐步剥离视频中的空间、时间冗余，再通过高效编码方式压缩剩余有效信息，成为历代视频编码标准的技术基石。

预测编码是框架的第一步，核心目标是利用视频的时空相关性减少冗余数据。帧内预测针对单帧内部的像素关联，通过分析同一帧中已编码区域的纹理特征，采用空间插值或多方向预测策略生成当前编码块的预测值，无需重复编码相似纹理区域，尤其适用于图像中纹理连续的区域。帧间预测则聚焦于连续视频帧之间的运动关联性，通过运动估计技术精准定位当前块在参考帧（已解码帧）中的对应位置，再通过运动补偿生成预测块，最终仅传输运动向量和预测残差（原始块与预测块的差异），大幅减少帧间重复信息的传输量，是实现高压缩比的关键环节。

预测之后的残差数据仍包含较多冗余，需通过变换与量化进一步压缩。变换过程将残差数据从像素域转换到频域，常用离散余弦变换（DCT）或其整数近似形式，利用自然图像残差的能量集中特性，将大部分能量汇聚到少数低频系数上，为后续压缩奠定基础。量化则是实现有损压缩的核心步骤，通过按预设步长降低频域系数的精度，尤其对人眼不敏感的高频系数进行更大幅度的精度压缩，从而显著减少数据量，量化步长可根据实际画质需求动态调整，平衡压缩效率与视觉效果。

熵编码作为框架的最后一步，对量化后的系数及预测过程中产生的辅助信息（如运动向量、预测模式标识等）进行无损压缩。它基于数据的概率分布特征，采用优化的编码模型减少编码冗余，常见的实现方式包括上下文自适应可变长度编码（CAVLC）和上下文自适应二进制算术编码（CABAC），其中CABAC 通过动态调整编码上下文，能更精准地匹配数据分布，实现更高的压缩效率。

此外，环路滤波作为重要的优化模块，在

解码端对重构图像进行处理，有效减少块效应、ringing 失真等编码 artifacts。典型的环路滤波技术包括去块滤波（DBF）和样本自适应偏移（SAO），前者针对性消除块编码带来的边界不连续问题，后者通过对样本值进行自适应偏移调整提升重构图像的准确性，两者共同提升解码图像质量，同时优化后续帧的预测精度，进一步增强整个编码框架的性能。

H.26x 系列（含 H.261、H.263、H.264/AVC、H.265/HEVC、H.266/VVC）是传统混合编码框架的标准化实现与迭代演进产物——H.26x 系列并非颠覆该框架，而是在其核心流程基础上，通过技术优化、功能扩展和细节革新，持续提升压缩效率与应用适配性。

H.261 作为该系列首个标准（1988 年），首次将混合编码框架落地为国际标准，定义了帧间运动补偿、8×8 DCT 变换、量化与熵编码的基础流程，为后续系列奠定架构基础。续 H.263 、 H.264/AVC 、 H.265/HEVC 、 H.266/VVC 均延续这一核心流程，在各环节的技术细节上进行迭代。

2.2 H.261 标准

H.261 作为 ITU-T 于 1988 年发布的首个面向实时视频通信的国际标准（ITU-T Rec. H.261, 1988），其核心定位是解决窄带综合业务数字网（N-ISDN）环境下视频会议与可视电话的传输需求，支持 p×64 kbit/s（p 取值 1-30）的灵活码率范围，为后续视频编码标准奠定了 “预测 - 变换量化 - 熵编码”的混合编码框架基础。该标准以 16×16 宏块为基本编码单元，采用帧间运动补偿预测、8×8 离散余弦变换（DCT）、变长编码（VLC）等核心技术，定义了 CIF（352×288 像素）与 QCIF（176×144 像素）两种标准输入格式，成功适配 NTSC 与 PAL 两种电视制式，解决了早期视频编码缺乏统一格式的兼容性难题（ITU-T Rec. H.261, 1993 修订版）。

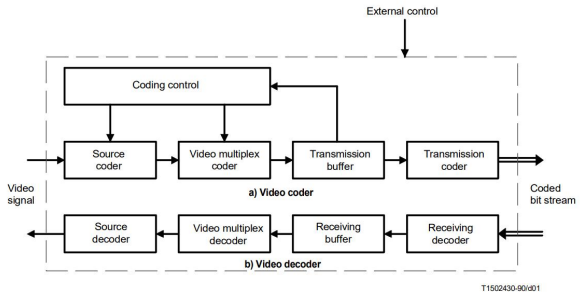


FIGURE 1/H.261
Outline block diagram of the video codec

在客观压缩效率指标上，相较于早期无标准方案，H.261 通过运动补偿复用、自适应扫描等技术的综合优化，亮度分量的编码误差均方根（RMS）显著降低，色度分量编码比特数也得到有效控制 —— 例如在 CIF 格式下，其对视频信号的冗余去除能力较传统模拟压缩技术提升约 30%，相同码率下的画质主观评分（满分 5 分）可达到 4-5 分的“可接受画质” 水平（ITU-T Rec. H.261, 1993）。在实时性与应用性能方面，H.261 在 384 kbit/s（p=6）码率下可实现 CIF 格式 15 帧 / 秒的实时编码与解码，配合可调节的帧速率（7.5/10/15 帧 / 秒），完全满足视频会议的实时交互需求；而在 1.92 Mbit/s 码率下，

其输出画质接近 VHS 水准，在当时的窄带网络环境中实现了画质与带宽的高效平衡（Richardson, I. E. G., 2010）。此外，H.261 还定义了假设参考解码器（HRD）与视频缓冲验证器（VBV），通过严格的缓冲器大小与码率约束（如 CIF 格式单帧编码比特数不超过 256 kbits，QCIF 格式不超过 64 kbits），确保不同实现方案的性能一致性，进一步巩固了其在早期实时视频通信领域的核心地位（ITU-T Rec. H.261, 1993）。

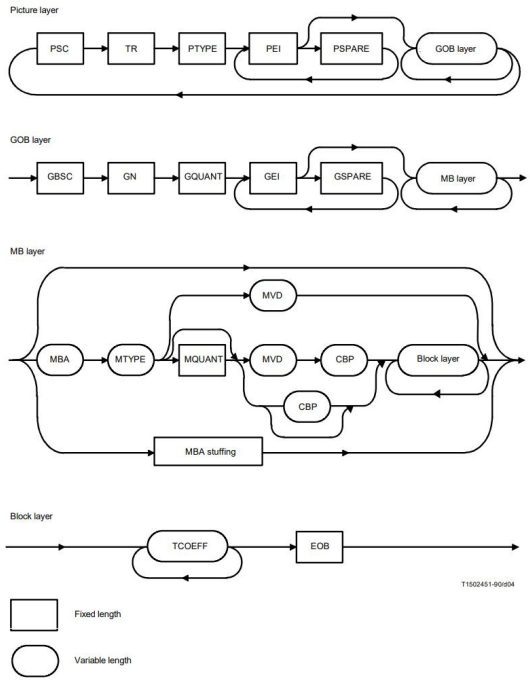


FIGURE 4/H.261
Syntax diagram for the video multiplex coder

2.3 H.262 标准

H.262（也被称为 MPEG-2 视频标准，对应 ISO/IEC 13818-2）是 ISO/IEC 与 ITU-T 联合制定的国际标准，作为面向数字存储媒体与数字电视传输的视频编码标准，其核心定位是解决 H.261 等早期标准在高清画质支持、多场景适配性等方面的不足，支持从标清到高清的多种分辨率，码率范围覆盖 1.5 Mbit/s 至 40 Mbit/s 以上，适配数字电视广播、DVD 存储、视频会议等多类应用场景。该标准继承了混合编码框架，以 16×16 宏块为基本编码单元，支持帧内编码（I 帧）、帧间预测编码（P 帧）与双向预测编码（B 帧），定义了亮度（Y）与色度（U、V）分量的 4:2:0、4:2:2 等多种采样比例，同时兼容隔行扫描与逐行扫描两种图像格式，为数字视频的标准化传输与存储奠定了基础（ISO/IEC 13818-2, 1994; Richardson, I. E. G., 2010）。

相较于 H.261，H.262 在相同主观画质下，码率开销降低约 30%-50%—— 例如在 CIF 格式、384 kbit/s 码率下，H.261 的峰值信噪比（PSNR）约为 32 dB，而 H.262 通过半像素运动补偿与 B 帧编码，PSNR 可提升至 35 dB 以上；在 720×576 标清格式下，H.262 以 4-6 Mbit/s 码率即可实现主观评分 4 分（满分 5 分）的清晰画质，满足数字电视广播需求（ISO/IEC 13818-4, 1996）。在实时性与应用性能方面，H.262 的编码延迟控制在 200 ms 以内，虽略高于 H.261 的 150 ms，但完全满足数字广播与存储回放的实时性要求；其支持的隔行扫描编码模式，完美适配传统电视信号传输，而 H.261 仅支持逐行扫描，难以兼容传统电视设备（Richardson, I. E. G., 2010）。此外，H.262 定义的假设参考解码器（HRD）与视频缓冲验证器（VBV），通过严格的缓冲器大小与码率约束，确保了

不同厂商实现方案的性能一致性，其多采样比例与多分辨率支持能力，使其成为数字视频时代的核心标准，广泛应用于全球数字电视广播、DVD 播放机等设备中（ISO/IEC 13818-2, 1994）。

2.4 H.263 标准

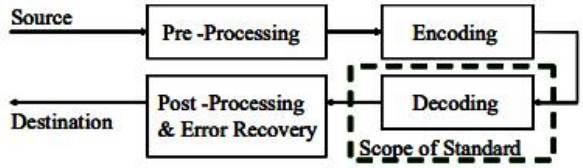
H.263 是 ITU-T 制定的低码率视频编码标准，最初于 1996 年发布，1998 年版本（又称 H.263+）进一步扩展了编码选项，旨在解决早期 H.261 标准在压缩效率、场景适配性和传输鲁棒性等方面的局限，成为多媒体通信领域的核心标准之一（Bormann et al., 1998; ITU-T Rec. H.263, 1996; ITU-T Rec. H.263, 1998）。该标准支持从 QCIF 到 CIF 乃至更高分辨率的视频格式，码率适配范围广泛，核心应用场景涵盖视频会议、可视电话、网络视频传输等低带宽需求场景，其 RTP 载荷格式通过 RFC 2429 标准化，确保了在互联网环境中的高效传输与互联互通（Bormann et al., 1998）。

H.263（含 1998 年 H.263 + 版本）作为 ITU-T 面向低码率多媒体通信的视频编码标准，与 ISO/IEC 13818-2 定义的 H.262（MPEG-2 Video）标准在设计目标与技术特性上形成明确区分，H.263 通过 RFC 2429 标准化其 RTP 载荷格式，优化互联网环境下的传输适配，而 H.262 聚焦广播与存储场景，强调高兼容性与多分辨率支持；相较于 H.262 采用的 8×8 DCT 变换与整数像素运动估计，H.263 引入半像素精度运动补偿与优化的 DCT 量化策略，H.263 + 更保留双向预测 B 帧并新增切片结构模式、独立段解码（ISD）、参考图像选择等机制，有效增强丢包环境下的传输鲁棒性，同时移除 H.262 中起始码的前两字节并通过载荷头部指示位提升传输效率（Bormann et al., 1998; Richardson, I. E. G., 2010）；在功能扩展性上，H.262 支持隔行 / 逐行扫描及多种轮廓与级别组合，适配从 QCIF 到 HDTV 的分辨率，而 H.263 突破 H.262 在低码率场景的局限，扩展了更多图像格式，H.263 + 新增时间、SNR、空间三种可扩展性模式，支持可变码率（VBR）与恒定码率（CBR）双模式，更适配窄带通信的动态带宽变化（Bormann et al., 1998）；低码率场景下 H.263 凭借更精细的压缩机制实现更高峰值信噪比（PSNR），画质损伤更小，且编码延迟控制在实时交互可接受范围，而 H.262 在高码率、高分辨率场景中更能保留细节，其假设参考解码器（HRD）与视频缓冲验证器（VBV）机制更适配稳定的广播信道传输。

2.5 H.264 标准

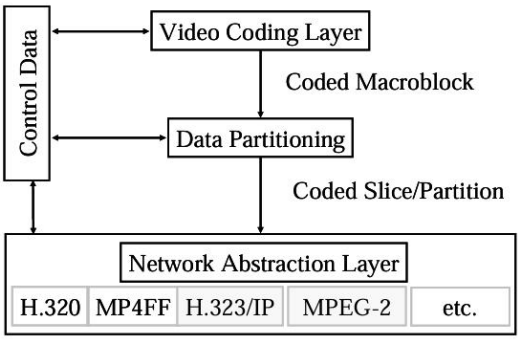
H.264/AVC 作为 ITU-T 与 ISO/IEC 联合制定的新一代视频编码标准（ITU-T Rec. H.264 | ISO/IEC 14496-10），于 2003 年正式发布，其核心设计目标是突破 H.263（含 1998 年 H.263 + 版本）的性能局限，实现更高压缩效率与更强的网络适配能力，涵盖广

播、存储、实时通信等多场景应用（Richardson, I. E. G., 2010; Wiegand et al., 2003）。该标准创新性地采用视频编码层（VCL）与网络抽象层（NAL）分离架构，VCL 聚焦高效视频内容压缩，NAL 则负责适配不同传输与存储场景，其多档次（Baseline、Main 等）设计进一步拓展了在不同终端与带宽条件下的适用性，成为后续视频编码技术的重要基础（Wiegand et al., 2003）。相较于 H.263，H.264/AVC 在预测机制、变换编码、熵编码等核心技术上实现全面革新，同时强化了网络传输鲁棒性，显著提升了整体性能（Bormann et al., 1998; Richardson, I. E. G., 2010）。



帧内预测上，H.263 + 仅支持简单空间预测，而 H.264/AVC 引入 4x4 与 16x16 块的多方向空间预测，提供 9 种 4x4 亮度预测模式与 4 种 16x16 亮度预测模式，精准匹配图像边缘纹理，大幅降低帧内冗余，同时针对色度分量设计专属预测机制，进一步优化平滑区域压缩效率（Richardson, I. E. G., 2010; Wiegand et al., 2003）。帧间预测领域，H.263 支持半像素精度运动补偿与有限块大小选择，而 H.264/AVC 采用 16x16 至 4x4 的可变块大小运动补偿，配合四分之一像素精度插值滤波，实现复杂运动场景的精准建模；同时支持多参考帧预测（最多 16 帧），解除 H.263 中 B 帧不可作为参考帧的限制，显著提升运动平缓区域与遮挡场景的编码效率（Wiegand et al., 2003; Bormann et al., 1998）。变换与熵编码环节，H.263 采用 8x8 DCT 变换，存在变换失配与块效应问题，而 H.264/AVC 采用 4x4 整数变换，避免编解码漂移，同时通过上下文自适应可变长编码（CAVLC）与上下文自适应二进制算术编码（CABAC）提升熵编码效率，CABAC 较 H.263 的固定 VLC 效率提升 5%-15%（Wiegand et al., 2003）。此外，H.264/AVC 新增环路去块滤波器，自适应消除块效应并反馈至预测环路，较 H.263 的简单后处理滤波大幅提升画质（Richardson, I. E. G., 2010）。H.264/AVC 的性能提升与网络适配能力已通过权威测试验证。压缩效率上，相同主观画质下，H.264/AVC 较 H.263 可降低 40%-60% 码率，例如 CIF 分辨率 30 帧 / 秒视频编码中，同等 PSNR 下 H.264/AVC 码率仅为 H.263 的一半左右（Richardson, I. E. G., 2010; Wiegand et al., 2003）。画质表现上，更精准的预测机制与去块滤波使块效应、模糊等 artifacts 大幅减少，低码率场景下优势尤为突出（Wiegand et al., 2003）。网络传输方面，H.263 + 通过切片结构、ISD 机制提升抗丢包能力，但缺乏统一网络抽象层，而 H.264/AVC 的 NAL 单元结构支持灵活切片分组（FMO）与任意切片排序（ASO），参

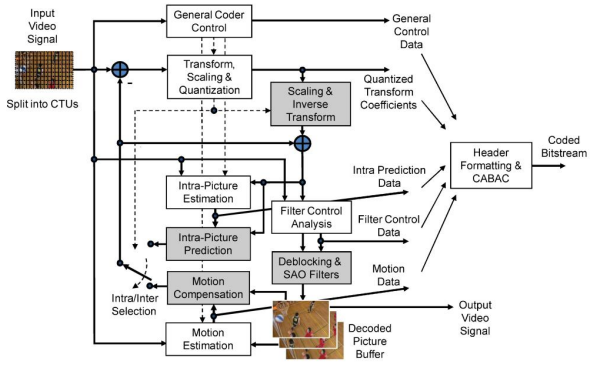
数集（SPS/PPS）设计降低关键信息丢失风险，支持冗余图像与数据分区传输，在 10% 丢包率环境下仍能保持可接受解码质量，避免 H.263 + 的帧内错误传播（Bormann et al., 1998; Wiegand et al., 2003）。此外，H.264/AVC 的低复杂度实现使其适配从移动终端到专业广播设备的各类硬件，满足不同场景的延迟与复杂度需求，成为数字视频领域的核心标准（Richardson, I. E. G., 2010）。



2.6 H.265 标准

H.265/HEVC（高效视频编码）作为 H.264/AVC 的继任标准（ITU-T Rec. H.265 | ISO/IEC 23008-2），于 2013 年正式发布，核心目标是在保持画质不变的前提下提升压缩效率，适配 4K/8K 等超高清场景，解决 H.264 在高分辨率编码时的带宽与存储压力（Sullivan et al., 2012; ITU-T Recommendation H.265, 2013）。相较于 H.264，H.265 的核心技术改进集中在多维度：编码单元方面，将 H.264 固定 16×16 宏块升级为 16×16、32×32、64×64 可变大小的编码树单元（CTU），通过四叉树动态划分适配不同纹理复杂度区域；预测机制上，帧内预测从 8 种方向扩展至 33 种，新增平面预测模式，帧间预测新增不对称运动分区（AMP）与融合模式（Merge Mode），配合 7 抽头 / 8 抽头插值滤波器提升复杂运动适配能力；环路滤波在去块滤波器基础上新增样本自适应偏移（SAO）滤波，进一步消除编码失真（Richardson, I. E. G., 2010; Sullivan et al., 2012）。

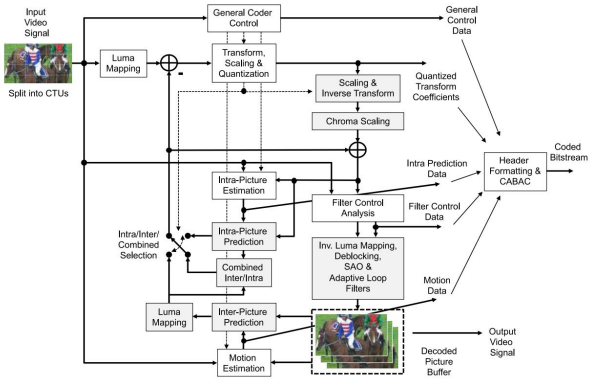
在相同主观画质下较 H.264 可降低 50% 码率，4K 视频编码中表现尤为突出，大幅降低了超高清视频的传输与存储成本（Richardson, I. E. G., 2010; Sullivan et al., 2012）。画质方面，更精细的预测与 SAO 滤波减少了块效应与模糊，低码率场景下细节保留能力更强；场景适配性上，支持最大 7680×4320（8K）分辨率与 10 位比特深度，适配超高清、多视角等新兴需求，同时引入瓦片（Tiles）、波前并行处理（WPP）等机制强化并行编码能力，适配多核硬件，兼顾移动端低复杂度与专业设备高性能需求（Sullivan et al., 2012; ITU-T Recommendation H.265, 2013）。



2.7 H.266 标准

H.266/VVC（多功能视频编码）作为 ITU-T H.266 与 ISO/IEC 23090-3 标准，相较于 H.265 以提升压缩效率为核心的设计，H.266 在技术架构上实现了多维度革新：块划分采用 QT+MTT 混合结构，将最大编码树单元（CTU）尺寸从 H.265 的 64×64 扩展至 128×128，同时支持亮度与色度独立分区（CST），更适配高分辨率视频的纹理差异特性（Bross et al., 2021; Sullivan et al., 2012）；帧内预测在 H.265 的 33 种方向基础上扩展至 93 种，新增矩阵基预测（MIP）与多参考线（MRL）技术，大幅提升空间预测精度，帧间预测则引入仿射运动模型、双向光流（BDOF）等工具，可精准捕捉缩放、旋转等复杂运动，弥补了 H.265 在非平移运动场景的适配短板（Bross et al., 2021）。此外，H.266 新增参考图片重采样（RPR）、子图片提取合并（BEAM）等高层功能，解决了 H.265 在分辨率动态切换、沉浸式视频局部解码等场景的技术局限（Bross et al., 2021）。

相同主观画质下，较 H.265 平均降低 50% 码率，UHD 分辨率视频的码率节省甚至可达 40% 以上，较 H.264 则实现 75% 的码率缩减（Bross et al., 2021; Richardson, 2010）。画质方面，通过自适应环路滤波（ALF）、亮度映射色度缩放（LMCS）等工具，H.266 在低码率场景下的块效应与色彩失真显著减少，尤其适配 HDR、广色域内容的编码需求，解决了 H.265 在高动态范围视频编码中细节保留不足的问题（Bross et al., 2021）。应用适配性上，H.266 支持 8K 及更大分辨率、360° 沉浸式视频、超低延迟流等新兴场景，其灵活的配置文件（如 Main 10、Main 10 4:4:4）可覆盖从移动端到专业广播设备的多元需求，而瓦片（Tiles）、波前并行处理（WPP）等机制的优化，进一步强化了并行编码能力，适配多核硬件环境（Bross et al., 2021; Sullivan et al., 2012）。早期实现验证显示，H.266 解码器可在主流硬件上实现 4K 60fps 实时解码，编码效率与实用性均较 H.265 实现质的提升。



2.8 总结

H.26x 系列视频编码标准的迭代历程是压缩效率、技术架构与应用适配性的持续进化：H.261 作为早期标准奠定了块基编码基础，H.262/H.263 优化了预测与变换机制以适配广播与低码率场景，H.264/AVC 通过混合编码架构、多参考帧等技术实现质的突破，成为广泛应用的经典标准；H.265/HEVC 进一步提升压缩效率，以 64×64 CTU、33 种

帧内预测方向等革新实现较 H.264 50% 的码率节省，适配 4K 等超高清需求；而 H.266/VVC 则在前者基础上全面升级，通过 QT+MTT 混合块划分、93 种帧内预测方向、仿射运动模型等技术，较 H.265 再降 50% 码率，同时新增 RPR、BEAM 等功能，拓展至 8K、360° 沉浸式视频等场景，并行处理与硬件适配性显著提升，完成了从基础压缩到多功能、全场景适配的跨越。

3. 基于深度学习的视频编码方式 [22]-[29]

基于深度学习的视频编码（又称神经视频压缩）是近年来兴起的、利用深度神经网络实现视频高效压缩的前沿技术。其核心思想在于：摒弃传统混合编码框架中大量依赖先验知识和手工设计的固定模块，转而使用端到端可训练的神经网络，直接从海量视频数据中学习最优的压缩表示与重建策略。相较于传统编码标准（如 H.264/AVC、H.265/HEVC）在既定架构上的优化，神经编码旨在从根本上重构视频压缩的流程与逻辑，从而超越现有标准的性能极限，更是探索面向网络传输的新一代智能压缩方法。

神经编码的发展并非一蹴而就，而是经历了从模仿传统框架到创新核心组件，再到构建全新范式的演进过程。这一演进背后，是神经网络强大的非线性拟合能力、概率建模能力与生成能力被逐步发掘并应用于解决压缩中的核心问题——如何更高效地去除时空冗余并编码必要信息。与传统标准需严格保证解码器一致性和硬件友好性不同，神经编码的研究更侧重于算法性能的边界探索，其模型灵活多样，但同时也面临计算复杂度高、标准化滞后等实用化挑战。以下将沿其技术发展脉络，分三个阶段详细阐述各阶段的核心框架、关键创新及其与传统编码的关联与对比。

3.1 框架奠基与概念验证 (2018-2020)

本阶段的研究核心是验证“用深度学习实现端到端视频编码”的可行性，其技术路径高度模仿传统混合编码的“预测-变换”架构，旨在用神经网络一对一地替代传统编码器中的各个功能模块，并证明这种“神经化”版本在性能上具备竞争力。

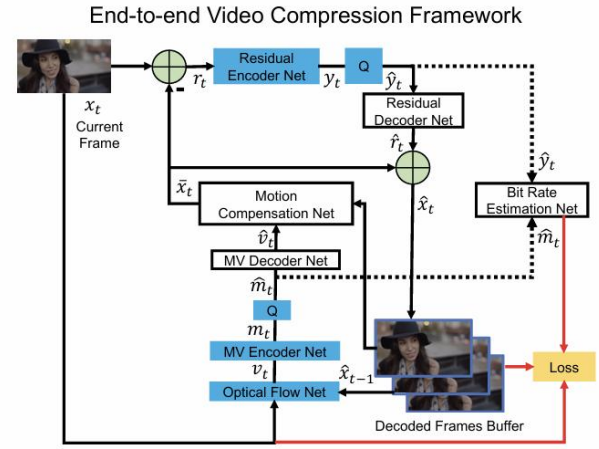
具体而言，研究主要围绕“运动补偿+残差编码”这一核心流程的神经实现展开，可视为对传统编码器（如 H.264, H.265）功能模块的一对一神经网络替代与映射。在运动估计与补偿环节，卷积神经网络（CNN）或专用光流估计网络（如 PWC-Net）取代了传统编码中基于块匹配的运动搜索算法，直接生成密集、连续的运动场（光流），随后通过可变形卷积这一可学习的操作实现运动补偿，从而突破了自 H.261 以来基于块的、仅能描述平移运动的传统模型局限，能够建模旋转、缩放等更复杂的局部形变。在残差处理部分，一个 CNN 自编码器被用于对预测后的残差进

行压缩，这一网络实质上融合并替代了传统流程中离散余弦变换（DCT）、量化以及熵编码等多个独立模块的功能，通过端到端训练从数据中学习一种非线性的、“内容感知”的最优变换方式，而非依赖于固定、线性的手工设计变换基。整个系统的训练核心在于一个全局可微的率失真损失函数 L=R+λ •D，其中码率 R 通过基于熵模型的比特估算得到，失真 D 则常用均方误差（MSE）或多尺度结构相似性（MS-SSIM）度量；通过调整拉格朗日乘子 λ，系统能够自动学习如何在码率与失真间进行权衡，这一机制在单一损失函数下实现了传统编码中需要复杂码率控制算法才能逼近的全局优化目标。

DVC 首次系统性地构建了这一完整框架，验证了端到端学习的可行性，其性能在 PSNR 指标上超越 H.264，在 MS-SSIM 指标上与 H.265 相当，为后续研究奠定了不可或缺的基线。紧随其后的 Scale-space Flow 工作则聚焦于改进运动表示，提出多尺度光流以显式建模运动不确定性，提升了处理大运动与遮挡场景的能力，体现了在奠基框架上对核心模块进行深度优化的初步努力。

总体而言，本阶段成功证明了神经视频编码的基本可行性，其框架本质上是传统编码的“神经网络重写版”，虽在压缩效率上尚未全面超越最新的传统标准 H.266/VVC，但已在特定感知指标上展现出潜力，并清晰地指明了性能瓶颈与后续优化的核心方向。

本阶段成功验证了深度学习视频编码的基本可行性，其框架可视为传统编码的“神经网络重写版”。虽然性能尚未全面超越最新传统标准（H.266/VVC），但在特定指标（如 MS-SSIM）上已展现出优势，并为后续研究提供了清晰的基线模型和优化方向。与传统编码相比，其最大区别在于所有模块的参数均从数据中学习得到，而非人工设计。



(b)

3.2 性能突破与核心创新（2020-2022）

在奠定基础框架并验证可行性之后，神经视频压缩研究进入了以性能突破为导向的密集创新阶段。本阶段的研究者不再满足于对传统架构的简单模仿，而是开始针对神经网络的内在特性，在运动表示、上下文建模、概率模型与整体架构等多个层面进行根本性创新，推动压缩效率迅猛提升，最终实现在客观率失真指标上逼近并超越最新传统标准 H.266/VVC。其中最关键的范式转变是从“预测编码”到“条件编码”的演进。

传统编码及早期神经编码均遵循“预测编

码”逻辑，即计算并编码当前帧与预测帧之间的像素差。而以 DCVC 为代表的工作则提出了革命性的“条件编码”范式：在特征域中，解码器不再试图精确重建残差，而是以一个从参考帧等信息源生成的、高维的“上下文特征”作为强条件，直接生成或合成出当前帧的表示。这跳出了“减法”去冗余的线性思维，转而利用神经网络的生成能力在丰富先验下“填补”信息，极大地增强了对纹理细节和高频内容的重建能力，成为性能跨越式提升的首要驱动力。与此相辅相成的是熵模型的深刻变革，传统编码依赖手工设计的上下文自适应二进制算术编码（CABAC），而神经编码则引入了基于学习的超先验网络与标准化流等生成式概率模型（如 CANF-VC），以建模潜在特征间更复杂的条件依赖关系，从而实现了比手工模型更精准的熵估计与更低的码率。

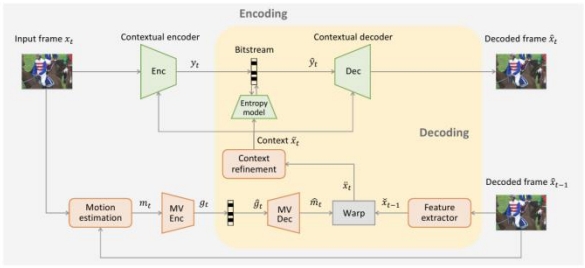
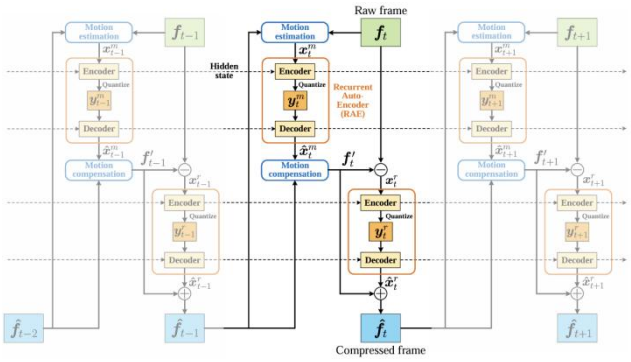


Figure 2: The framework of our DCVC.

在时序建模方面，研究也超越了早期单参考帧的局限，RLVC 通过引入循环神经网络（RNN）在编码器与解码器中建立记忆状态，有效地利用了长时帧间依赖；而 Deep Hierarchical Video Compression 则进一步设计了分层概率预测架构，通过多尺度变分自编码器（VAE）来表征视频，不仅提升了压缩效率，还显著降低了内存与计算开销，并首次在神经编码中实现了对网络传输极为重要的渐进解码功能。

这些创新共同将神经视频编码推向了新的高度，实验表明，顶尖方法在多种测试序列上已能实现对优化后 H.266/VVC 编码器的显著码率节省，标志着神经编码从“有潜力”正式迈入了“有竞争力”的新阶段，并开始形成一套独立于传统混合编码的、以数据驱动与条件生成为核心的新设计哲学。



3.3 前沿探索与新范式 (2022 至今)

当前，神经视频压缩的研究前沿已进入“无人区”，其焦点不再局限于在现有评价体系下超越传统编码，而是转向探索可能彻底颠覆视频表示与压缩逻辑的全新范式。这些探索旨在重新回答“什么是视频最本质的紧凑表示”以及“如何为人类视觉或机器感知进行极致压缩”等根本性问题，呈现出多元化、跨领域和基础性的特点。一个显著趋势是 Transformer 架构的全面引入与深化。

传统编码及前期神经编码主要依赖 CNN，

而如ST-XCT等工作将3D Transformer与创新的交叉协方差注意力机制深度集成到特征提取、帧重建与熵建模等所有关键环节。Transformer的全局注意力机制能够直接建模视频中长程的空间与时域依赖，这种能力是局部操作的CNN和传统基于块的编码所欠缺的，为学习更高效的时空联合表示开辟了新路径。另一条更具颠覆性的路径是隐式神经表示与生成式压缩的探索。隐式表示研究（如SNP）将视频视为连续的时空信号，用隐式神经网络（如多层感知机MLP）的参数来表征；压缩过程转化为优化并传输这些网络权重。这完全摒弃了“帧”、“块”、“像素”等离散概念，在理论上为任意分辨率、任意帧率的通用压缩提供了统一方案，是对传统帧基编码框架的根本性挑战。

与此同时，基于扩散模型的生成式压缩在极低码率领域异军突起，其核心思想是放弃对像素的精确重建，转而传输高度压缩的语义或运动信息，在解码端利用扩散模型等强大生成器“想象”出视觉质量优异的画面。这追求的是视觉感知质量的最优化，而非像素级的保真度，标志着压缩目标从经典的“率失真优化”向“率感知优化”的深刻转变。此外，诸如VSP等研究从更上游的视频表示学习入手，通过将视频序列建模为随机过程并利用过程对比学习来捕获其平稳的动态规律，所获得的紧凑表示天然适合于压缩任务。这类工作为构建能够理解视频时空本质的“基础压缩模型”提供了前瞻性思路。

综上，本阶段的研究正在绘制视频压缩技术的未来蓝图，描绘了一个可能完全脱离传统混合编码框架、深度融合生成式人工智能、并能自适应不同终端与任务需求的智能压缩新时代。这与H.26x系列在既定轨道上的稳健演进形成了鲜明对比，代表了两种截然不同的技术发展哲学与可能性边界。

4. 结尾

综上所述，视频编解码技术的发展历程是一部从经典工程优化迈向数据驱动智能创新的演进史。传统混合编码框架以H.26x系列为代表，通过数十年持续迭代，在压缩效率、网络适配性与硬件兼容性上达到了成熟的高度，支撑起了从标清到8K超高清、从广播电视到实时通信的广泛应用生态。然而，随着视频数据量的爆炸式增长与应用场景的多元化，传统方法在性能提升上逐渐面临边际效应递减的挑战。

与此同时，基于深度学习的视频编码技术以其数据驱动、端到端优化的特点，正开辟一条全新的技术路径。从最初对传统架构的神经化重写，到条件编码、概率建模与Transformer等核心创新，神经编码不仅在客观指标上逐步逼近乃至超越H.266/VVC等最新传统标准，更在感知质量、自适应压缩与语义保持等方面展现出独特优势。当前，隐式神经表示、扩散模型生

成压缩等前沿探索，正在重新定义视频的表示与重建逻辑，推动编码技术从“保真重建”向“感知优化”演进，为元宇宙、全息通信、机器视觉等未来应用奠定基础。

尽管神经编码在标准化、计算复杂度与硬件部署等方面仍面临挑战，但其代表的技术范式转移已不可逆转。未来，视频编解码技术将深度融合人工智能、计算机视觉与通信理论，走向更智能、更自适应、更跨场景的新阶段，持续赋能数字视觉生态的进化与革新。

参考文献：

[1] ITU-T. (1988). Recommendation H.261: Video codec for audiovisual services at n × 384 kbit/s. International Telecommunication Union.

[2] ITU-T. (1993). Recommendation H.261: Video codec for audiovisual services at p x 64 kbit/s. International Telecommunication Union.

[3] ISO/IEC 11172-5 (1994). *Information technology - Generic coding of moving pictures and associated audio information - Technical Report*.

[4] Richardson, I. E. G. (2010). *The H.264 Advanced Video Compression Standard* (2nd ed.). Wiley.

[5] Richardson, I. E. G. (2010). *The H.264 Advanced Video Compression Standard* (2nd ed.). Wiley.

[6] ISO/IEC DTR 11172-5 (1994). *Information technology - Generic coding of moving pictures and associated audio information - Technical Report*.

[7] ISO/IEC 13818-4 (1996). *Information technology - Generic coding of moving pictures and associated audio information - Part 4: Conformance*.

[8] Bormann, C., et al. (1998). RTP Payload Format for the 1998 Version of ITU-T Rec. H.263 Video (H.263+). *IETF RFC 2429*.

[9] ITU-T Recommendation H.261 (1993). *Coding of moving pictures and associated audio for visual telephony services at p×64 kbit/s* (Revised).

[10] ITU-T Recommendation H.263 (1996). *Video Coding for Low Bit Rate Communication*.

[11] ITU-T Recommendation H.263 (1998). *Video Coding for Low Bit Rate Communication*.

[12] Richardson, I. E. G. (2010). *The H.264 Advanced Video Compression Standard* (2nd ed.). Wiley.

[13] ISO/IEC 14496-10 (2003). *Information technology - Coding of audio-visual objects - Part 10: Advanced Video Coding (AVC)*.

[14] Wiegand, T., Sullivan, G. J., Bjontegaard, G., & Luthra, A. (2003). Overview of the H.264/AVC video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7), 560-576.

[15] ITU-T Recommendation H.263 (1998). *Video Coding for Low Bit Rate Communication*.

[16] Richardson, I. E. G. (2010). *The H.264 Advanced Video Compression Standard* (2nd ed.). Wiley.

[17] Sullivan, G. J., Ohm, J., Han, W., & Wiegand, T. (2012). Overview of the High Efficiency Video Coding (HEVC) Standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(12), 1649-1668.

[18]ISO/IEC 23008-3（2022）.Information technology - High efficiency coding and media delivery in heterogeneous environments - Part 3: 3D audio

[19] Bross, B., Wang, Y.-K., Ye, Y., et al. (2021). Overview of the Versatile Video Coding (VVC) Standard and Its Applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10), 3736-3764.

[20] Sullivan, G. J., Ohm, J.-R., Han, W.-J., & Wiegand, T. (2012). Overview of the High Efficiency Video Coding (HEVC) Standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(12), 1649-1668.

[21] Richardson, I. E. G. (2010). *The H.264 Advanced Video Compression Standard* (2nd ed.). Wiley.

[22] Lu, Guo, et al. "Dvc: An end-to-end deep video compression framework." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[23] Agustsson, Eirikur, et al. "Scale-space flow for end-to-end optimized video compression." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[24] Yang, Ren, et al. "Learning for video compression with recurrent auto-encoder and recurrent probability model." IEEE Journal of Selected Topics in Signal Processing 15.2 (2020): 388-401.
[25] Li, Jiahao, Bin Li, and Yan Lu. "Deep contextual video compression." Advances in Neural Information Processing Systems 34 (2021): 18114-18125.
[26] Ho, Yung-Han, et al. "Canf-vc: Conditional augmented normalizing flows for video compression." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.
[27] Lu, Ming, et al. "Deep hierarchical video compression." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 8. 2024.
[28] Chen, Zhenghao, et al. "Neural video compression with spatio-temporal cross-covariance transformers." Proceedings of the 31st ACM International Conference on Multimedia. 2023.
[29] Zhang, Heng, et al. "Modeling video as stochastic processes for fine-grained video representation learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.