电子测量技术
Electronic Measurement Technology
ISSN 1002-7300,CN 11-2175/TN

《电子测量技术》网络首发论文

题目：基于深度学习的 VVC 快速帧内模式决策
作者：施金诚，杨静
网络首发日期：2022-02-26
引用格式：施金诚，杨静．基于深度学习的 VVC 快速帧内模式决策[J/OL]．电子测量技术. https://kns.cnki.net/kcms/detail/11.2175.TN.20220225.1450.026.html

网络首发时间：2022-02-26 13:13:03
网络首发地址：https://kns.cnki.net/kcms/detail/11.2175.TN.20220225.1450.026.html

基于深度学习的 VVC 快速帧内模式决策

施金诚　杨静
（上海海事大学 信息工程学院，上海 201306）

摘　要：新一代视频压缩标准（H.266/VVC）在帧内预测中提供了 67 种预测模式，这使得编码效率得到大大提升，但同时也带来了极高的计算复杂度。本文提出了一种基于深度学习的帧内模式决策快速算法。首先针对编码块尺寸划分后块的大小形状不同的问题，对提取的亮度块进行预处理，并通过随机剪裁、重采样和卷积神经网络（CNN）上采样的方式，保证块的大小和质量。然后精心设计了 CNN 架构来降低帧内预测复杂度，并提出将当前编码块、相邻参考块以及残差块三者作为网络的输入，把率失真决策过程转换为分类问题，减少不必要的模式遍历。为训练所提出的深度学习网络，本文针对 H.266 的特点建立了模式决策数据集。实验结果表明，文章提出的算法与 VTM10.0 相比，编码时间平均降低了 39.56%~43.45%，有效的降低了编码的计算复杂度，同时率失真性能基本保持不变，与最新参考文献相比综合性能也有所提升。

关键词： H.266/VVC；帧内预测；模式决策；深度学习；上采样
中图分类号： TP919.81　**文献标识码：** A　**国家标准学科分类代码：** 510.60

VVC fast intra mode decision based on deep learning

SHI Jincheng　YANG Jing
(College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China)

Abstract: The new generation of video compression standard (H.266/VVC) provides 67 prediction modes in intra-frame prediction, which greatly improves the coding efficiency, but also brings extremely high computational complexity. This paper proposes a fast algorithm for intra-mode decision-making based on deep learning. First, for the problem of the size and shape of the block after the size of the coding block is divided, the extracted brightness block is preprocessed, and the block size and quality are guaranteed through random cropping, resampling, and convolutional neural network (CNN) upsampling. . Then the CNN architecture is carefully designed to reduce the complexity of intra prediction, and it is proposed to use the current coding block, the adjacent reference block and the residual block as the input of the network to convert the rate-distortion decision-making process into a classification problem and reduce unnecessary Pattern traversal. In order to train the proposed deep learning network, this paper establishes a model decision data set based on the characteristics of H.266. Experimental results show that compared with VTM10.0, the algorithm proposed in the article reduces the coding time by 39.56%~43.45% on average, which effectively reduces the computational complexity of coding, while the rate-distortion performance remains basically unchanged, which is comparable to the latest references. The overall performance has also been improved.

Keywords: H.266/VVC; intra prediction; mode decision; deep learning; upsampling

0　引　言

随着短视频、4K、8K 高清视频以及 VR 视频的快速发展，原本的 H.265/HEVC 已经无法满足人们的需要，因此成立了联合视频专家团队（JVET），旨在探索可提供超越 H.265 / HEVC 压缩性能的视频编码标准，在 2018 年首次定义了新一代视频编码标准多功能视频编码（即 VVC）并在同年发布了 VVC 测试模型（VTM1.0）[1]，最终在 2020 年宣布完成多功能视频编码（VVC）标准，截至本文撰写已经发布最新 VVC 测试模型 VTM-12.0。与上一代视频编码标准（HEVC）相比，在相同质量下可以提高约 40%的编码效率[2]，但 VVC 复杂度约是 HEVC 的 12 倍[3]。

为了减少 VVC 的帧内预测复杂度，已经提出很多优化的方法。在传统算法上杨浩等[4]提出将四叉树加多类型树划分（QTMT）建模为一个树状形的二进制分类问题，并加入统计学习的方法来决策编码单元（CU）大小；张秋文等[5]提出将随机森林分类模型（RFC）用在了快速 CU 划分，同时提出了基于纹理区域特征的快速帧内预测模式优化算法。

基金项目：国家自然科学基金资助项目（61902239）。

近几年随着深度学习的发展和应用，也开始被应用于视频编码。一般来说，基于深度学习的视频编码可以分为两类，第一类神经网络将未压缩的视频作为输入，直接输出压缩后的比特流，这种神经网络模型称为端到端模型，不在本文的讨论范围；第二类方法仍然遵循传统的基于块的混合视频编码框架，将神经网络集成到框架中以提高特定模块的性能。对于帧间预测，赵正辉等[6]提出使用卷积神经网络模型(CNN)将两个预测块组合起来，以增强双向帧间预测；Lee J K 等[7]提出使用神经网络预测来生成用于运动估计和补偿的虚拟参考帧；对于熵编码，马长岳等[8]提出了使用 CNN 来预测帧内预测相关语法元素的概率分布；对于环路滤波，王明泽等[9]提出了一种基于注意力的环路滤波器来代替原本的滤波器，它可以同时处理亮度和色度分量。

除此之外，对于改进帧内预测，也可以分为直接通过神经网络生成预测块和降低帧内预测复杂度这两种方式，对于前者，李家豪等[10]提出了一个全连接网络来学习从相邻多条参考线到帧内预测块的端到端映射；胡月雨等[11]提出了一个渐进空间循环神经网络(PS-RNN)来进行帧内预测，PS-RNN 以根据相邻内容逐步生成预测。对于后者，李天一等[12]提出了一种具有早退机制的多阶段退出 CNN 模型（MSE-CNN）和一种自适应损失函数，以根据多阶段的灵活 QTMT 结构来确定 CU 分区和最小化 RD 成本，这大大加快帧内模式 VVC 的编码过程；Tissier A 等[13]提出通过 CNN 训练来预测每个 4*4 划分的概率，进而降低帧内预测复杂度；Lin T L 等[14]提出了一种基于卷积神经网络的帧内模式决策方法，它在低分率类的视频中有着良好的表现，但在面对复杂的情况时，尤其是在高分辨类中，与其余的快速算法相比，没有明显的优势。

虽然之前基于深度学习的帧内预测方法已经取得了显著的性能，但仍然有很大改进潜力。以前的方法多是选择直接对当前块进行模式预测，仅仅考虑当前块的纹理信息，如文献[14]，这可以为大多数块生成准确的模式预测，然而，对于一些包含复杂纹理的复杂块，如混合多方向纹理或循环模式，尤其是在 H.266 标准下，仅使用当前块不足以产生准确的预测。为了进一步提高这些复杂块的预测精度，本文提出将当前块的 L 形相邻块以及当前块的预测残差也考虑进去，相邻块可以带来更多的局部上下文信息，而预测残差包括更多的压缩信息，同时残差块的有效性已经在后处理中的得到了证明。

同时针对 H.266 的特点，CU 块因多类型树划分导致的不规则矩形的问题，有的提出了简单的重采样的方式，但这大大破坏了原本图像块的纹理特征，而在文献[14]中，单一的选择了 16*16 的块作为输入，在文献[15]中则针对所有 CU 大小训练许多网络[15]，这又限制了网络的利用率。本文想设计一个尽可能满足所有 CU 的算法模型，同时也希望保证 CU 块的质量，因此设计了一个基于 CNN 上采样的预处理框架。综上本文的主要贡献如下：

1）提出了一种新的基于深度学习的帧内模式预测网络。除了参考当前编码块，还利用了相邻的 L 形参考块和相应的残差分量。

2）网络架构经过精心设计，可以同时使用三个输入组件，同时应用了通道注意力机制来有效地组合来自不同输入的特征。

3）针对 H.266 标准，设计了一个基于 CNN 上采样的数据预处理框架。

1 基于深度学习的帧内模式决策算法

在本节中，将介绍本文算法的细节，包括总体框架、数据预处理、决策网络架构以及数据集的建立。

1.1 总体框架

本文统计了五种不同序列在三种量化参数（QP）（20，25，35）下最终的帧内模式分布情况，如表 1 所示，当 QP 为 35 时，平均有 54.04%和 6.13%的 CU 选择“Planar”和“DC”模式作为最佳模式，也就是说大约 60%的 CU 选择非定向模式作为最佳模式，对于平坦区域较多的图像序列则更明显，如“FoodMarket”和“BasketballDrive”。随着 QP 的降低，量化的越来越精细，发现其定向角度模式的比例开始增加，当 QP=20 时，如图 1 所示，在具有丰富纹理序列“PartyScene”中，达到 72.73%的 CU 选择定向角度模式作为最佳模式。

表 1 帧内各模式所占比例

序列	QP	Planar (%)	DC (%)	2-66 (%)
PartyScene	20	18.53	8.74	72.73
	25	27.31	6.78	65.91
	35	44.27	2.52	53.21
CampfireParty	20	23.87	8.25	67.88
	25	29.81	9.75	60.44
	35	47.37	7.13	45.50
CatRobot	20	26.13	14.69	59.18
	25	45.21	12.46	42.33
	35	52.75	9.87	37.38
FoodMarket	20	27.06	9.72	63.22
	25	47.14	5.52	47.34
	35	59.67	4.32	36.01

	20	40.93	7.28	51.79
BasketballDrive	25	51.75	13.21	35.04
	35	66.12	6.81	27.07
	20	27.30	9.74	62.96
Average	25	40.24	9.54	50.21
	35	54.04	6.13	39.83

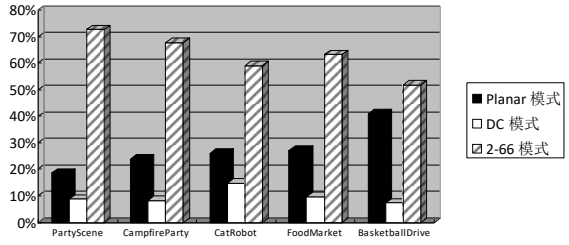

图 1 QP 为 20 时帧内模式分布

本文希望利用这些占比概率，进行网络的训练优化，对所选择的帧内预测模式的决策公式化为分类问题，从而避免原本复杂的 RDO 过程，来减少计算复杂度。

以下是本文的总体框架，如图 2 所示，经过块划分的 CU 块首先经过本文设计的预处理，之后进行本文设计神经网络的预测，以选出最佳帧内模式。

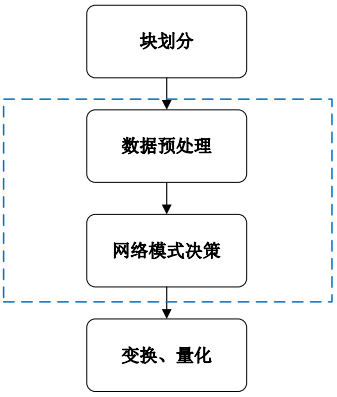

图 2 算法总体框架

1.2 数据预处理

本文统计了不同大小块的编码单元在数据集中所占的比例，如表 2 所示，发现较小的块所占的比例很高，例如 4*4、4*8、8*4 以及 8*8，但是它们所覆盖的面积很少，相反较大的块（64*64、32*32），虽然所占比例比较低，但是它们的覆盖区域更多，这是很好理解的，对于较大的块，即使块数量较少，覆盖面积也可以更大。

表 2 不同大小块的编码单元比例

CU 尺寸	A1	A2	B	C	D	E	平均	平均面积
64×64	0.49%	3.73%	0.06%	0.00%	0.64%	0.12%	0.84%	3358.72
32×32	0.62%	7.04%	1.39%	0.24%	0.36%	0.91%	1.76%	1781.76
32×16	0.29%	8.69%	4.08%	0.28%	0.38%	0.99%	2.45%	1254.4
16×32	0.53%	6.58%	0.88%	0.30%	0.48%	5.52%	2.38%	3706.88
32×8	0.15%	3.14%	8.80%	0.13%	0.77%	0.99%	2.33%	596.48
8×32	0.46%	1.65%	0.77%	0.16%	0.69%	1.66%	0.90%	230.4
16×16	3.19%	31.75%	4.88%	3.86%	1.67%	5.59%	8.49%	2168.32
32×4	0.08%	0.92%	22.36%	0.37%	2.35%	1.10%	4.53%	579.84
4×32	0.45%	0.28%	0.78%	0.04%	0.88%	2.36%	0.80%	102.4
16×8	2.86%	14.64%	9.58%	5.51%	5.32%	5.52%	7.24%	926.72
8×16	4.26%	12.93%	2.96%	3.87%	1.99%	9.41%	5.90%	755.2
16×4	2.30%	0.59%	6.24%	9.66%	6.62%	2.82%	4.71%	301.44
4×16	4.27%	0.33%	0.94%	1.42%	2.19%	7.45%	2.77%	177.28
8×8	15.14%	5.22%	14.87%	15.15%	10.05%	15.14%	12.60%	806.4
8×4	19.98%	1.30%	12.59%	25.97%	18.11%	11.23%	14.86%	475.52
4×8	26.41%	1.12%	4.67%	11.74%	15.22%	22.41%	13.60%	435.2
4×4	18.73%	0.10%	4.14%	21.31%	32.27%	11.08%	14.61%	234.92

考虑到在输入网络的图像块不宜太小，CU 太小在卷积过程中可能丢失一些信息，同时在做网络预测时也要尽可能的保留各个 CU 块的特征，包括大小，因此选择只对较小的 CU 块进行上采样处理，如 4*4、8*8，以符合神经网络的输入需要。以往常用的上采样方法包括：最邻近元法、双线性插值上采样等，它们都被广泛运用于图像处理中，但是缺点也很明显，当遇到复杂内容时，基于固定插值滤波器的上采样表现就不符预期，面对这样的问题，本文提出了基于 CNN 的上采样模型，具体处理过程如图 3 所示。

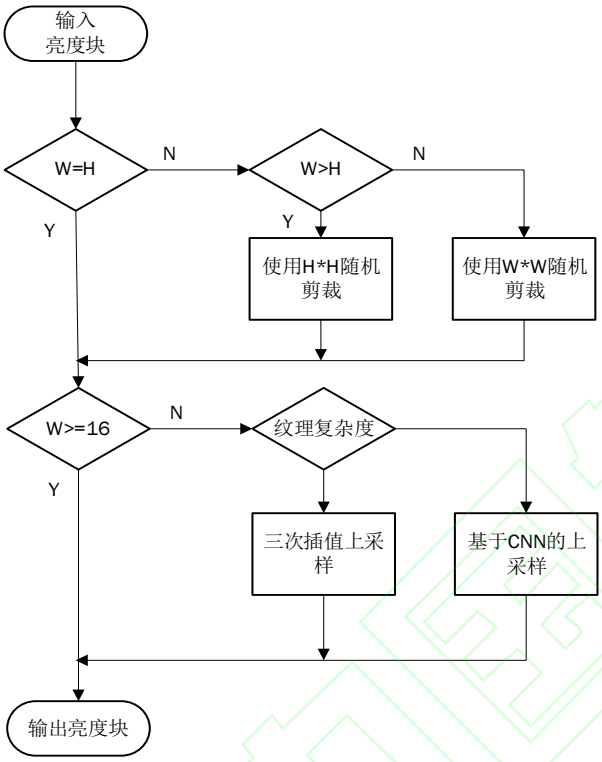

图 3 预处理流程图

通过块的纹理复杂度设置阈值，如式（1）所示，其中 Z 表示灰度，$P(Z_i)$ 为相应的直方图，L 是不同灰度级的数量，m 是 Z 的均值，如式（2）所示。对于相对简单的 CU 块直接采用双三次插值上采样，而对于相对较复杂的块则采用设计的 CNN 上采样的方式，这样设置阈值的好处可以避免无意义的上采样所带来的整个编码时间的增加。本文希望采用这种预处理方式，在满足网络训练的同时，尽可能保留原本图像的纹理方向等信息，以提高预测网络性能。

$$\mu(Z)=\sum_{0}^{L-1}(Z_i-m)^2\,P(Z_i) \qquad (1)$$

$$m=\sum_{0}^{L-1}Z_iP(Z_i) \qquad (2)$$

为了使用浅层的网络实现高质量的图像上采样，本文借鉴了以往的论文中的一些要素[16]，包括特征提取、反卷积以及多尺度融合[17]，具体网络模型如图 4 所示。

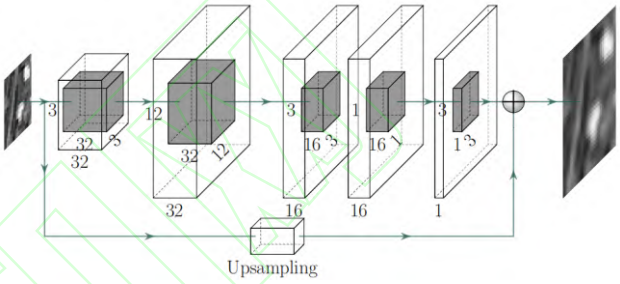

图 4 基于 CNN 的上采样模型

1.3 决策网络架构

本文提出的整个网络如图 5 所示，采用三个输入部分，包括当前编码 CU 块和 L 形相邻参考线，另一方面，相应的残差块分量用作第三个输入，因为残差快包含有关纹理特征的额外信息，为了同时利用这三个输入，本文精心设计了网络架构以生成更好的预测。

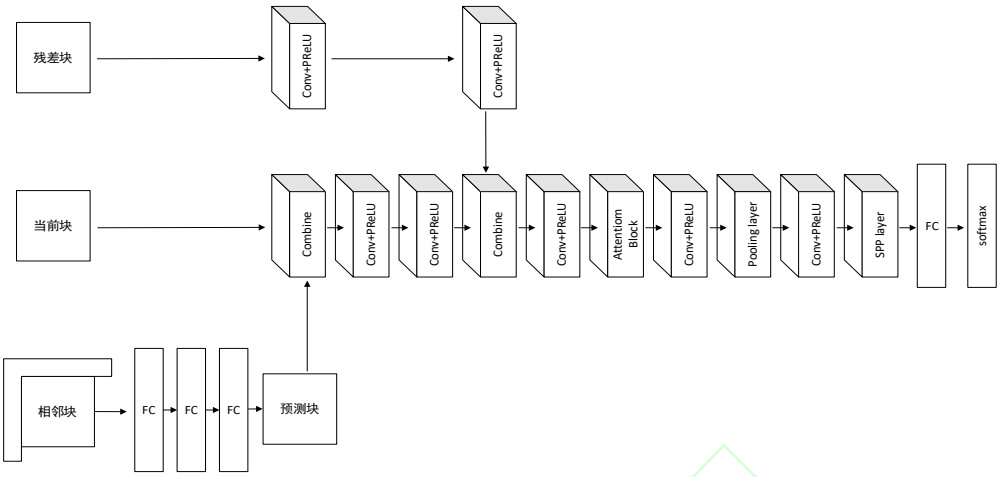

图 5 决策网络架构

其中当前编码 CU 块仍然被用作主要推理源，为了提供更多的局部上下文信息，我们使用全连接层(FC)结构从相邻 L 型参考线中提取特征，对于 N*N 大小的块，FC 结构的输入是相邻的 4N+1 块，如图 6 所示，这样可以提供更多的局部相关性信息。

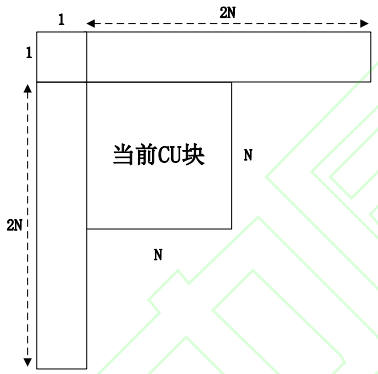

图 6 L 型参考线

同时将残差分量引入作为补充信息有助于减少噪声，提高我们的预测。我们设计了一个简单结构，包括两个卷积层，通过一个激活函数来提取相应的特征，提取的残差特征与从其他两个输入中提取的特征相结合，形成后续的输入。

为了充分利用这三个输入，本文进一步选择使用 channel-wise attention 来组合这些特征图，具体来说，Squeeze-and-Excitation Block[18]用作通道注意力单元，特征进行重新校准通过 SE 块中名为 Squeeze 和 Excitation 的两个步骤执行。先将原本的特征图通过挤压步骤得到 1*1*C 的融合特征，之后将其用作 Excitation 步骤的输入，以推导出新的通道权重，这些通道权重用于缩放原始输入特征图并生成得到新加权特征图，将原本提出的三个输入中提取的连接特征图进行相比，使用通道注意单元可以以更有效的方式得到融合特征图[19]。这些不同种类的特征图根据它们对帧内预测的相对重要性进行重新组合，这些聚合后的特征图直接输入到后续的卷积层中以生成最终的帧内模式。

在最后的池化层中，本文参考了 SPP-Net[20]，SPP-Net 特有的空间金字塔池化层非常适合用于帧内模式预测中，可以解决原本 CU 块尺寸不一的问题，如图所示。我们希望通过它从不同尺度的卷积层输出上提取特征，并且映射到尺寸固定的完全连接层上。在本文的网络中，每个卷积层都包含 64 个 3*3 的卷积核，我们将参数 ReLU (PReLU) 作为激活函数，其中需要在训练阶段学习尺度参数，我们还采用残差学习策略来实现更快的收敛。

在损失函数上，本文采用交叉熵作为网络的损失函数，如式（3）所示，其中 m 为当前 batch 的样本数，n 为标签数，这里 m 取 64，n 取 67 代表 67 种帧内预测模式，y_{ji} 为真实概率分布，$\hat{y}_{ji}$ 为预测的概率分布。

$$Loss = -\frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n} y_{ji} log_{e}\left(y_{ji}\right) \qquad (3)$$

1.4 数据集的建立

CTU 新的分区结构是 VVC 标准的主要贡献之一，VVC 在 HEVC 四叉树的基础上，加入了多类型树划分，包括新的竖直二叉划分 (SPLIT_BT_VER)、水平二叉划分 (SPLIT_BT_HOR)、竖直三叉划分(SPLIT_TT_VER)、水平三叉划分(SPLIT_TT_HOR)，这种划分方式会使得划分结果更加灵活，但也带来了新的问题，过往针对 H.265 所建立的数据集将不再适用 H.266，因此本文首先针对 H.266 所述的特点，为 VVC 帧内预测建立了一个所需要的数据集。选

择了多个不同种类、不同分辨率的视频序列，这确保了学习预测 CU 模式决策训练数据的充足和多样性。最后本文的数据集来自 30 个不同种的视频序列（如表 3 所示），这 30 个序列主要来源于 JCT-VC（Joint Collaborative Team on Video Coding）会议[21]同时包含了 5 种不同的分辨率序列：416*240、768*512、1536*1024、1920*1080 和 2560*1600。所有视频序列均以四种量化参数 QP（22，27，32，37）通过配置文件“encoder_intra_main.cfg”在全帧内（AI）下进行编码，参考软件为 VTM10.0，之后在解码端提取每个亮度块的最佳模式、以及其 L 形参考像素和相应残差分量，用作标签。

表 3 数据集来源信息表

分辨率	序列数	总帧数
416*240	8	240
768*512	7	210
1536*1024	7	140
1920*1080	6	150
2560*1600	2	60
总数	30	800

在本文建立的数据集中，将其中的 90%用作预测网络的训练集，另外 5%用作验证集，5%用作测试集。

2 实验结果与分析

首先在 GPU 服务器 GTX 1080 Ti 上进行本文卷积神经网络的训练，深度学习框架使用的是 Tensorflow2.10，Bach size 为 64，优化函数为 Adam， learnling rate 为 0.001。之后将本文训练好的模型嵌入 VTM10.0 进行编码测试，测试环境硬件配置为 Intel(R) Core(TM) i5-7200 CPU，主频为 3.20GHZ，内存为 8GB，操作系统为 64 位 Windows10。

为了验证本文所提出的模式预测算法的准确率，本文首先在 VTM10.0 上对不同分辨率不同类的五种序列在 AI 模式下进行编码，并记录下对应的最佳模式，之后将这五种序列用于本文提出的算法，同时输出其的最佳模式，以原编码器为基准，比较两者最佳模式的重合率，重合率 H 定义如式（4）所示，其中 N_{hit} 是本文算法预测帧内模式与原 VTM 中一样的 CU 个数，N_{total} 则为总的 CU 个数。

$$Hitratio = \frac{N_{hit}}{N_{total}} \quad (4)$$

较高的准确率可确保较高的预测有效性，同时带来更好的 RD 性能，从图 7 中可以发现，所有测试序列的准确率约为 90%，此外准确率随着 QP 的增加而略有增高。

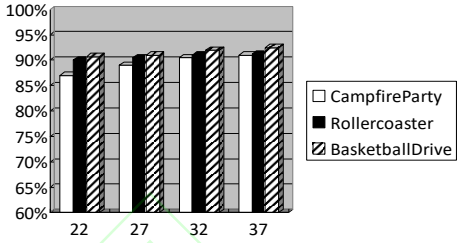

图 7 帧内预测模式准确率

之后对所提出的算法做更进一步的评估，将本文提出的算法嵌入到 VTM10.0 中，测试集选择 JVET 指定的标准 CTC 序列，总共 18 个视频序列，五个类别：A（3840x2160），B（1920x1080），C（832x480），D（416x240），E（1280x720）。对选取的所有序列的前 50 帧使用 AI 全帧内编码配置在四种 QP（22、27、32 和 37）下进行编码测试。编码性能采用 Bjontegaard Delta Bit 比特率（BD-BR）和编码减少的时间ΔT 进行评估。BD-BR 表示相同图像质量下，参考编码器和测试编码器之间的比特率平均差，当 BD-BR 得出为负时，表示节省了码率，说明当前测试编码器比参考编码器更有效[22]。ΔT 则表示编码时间的变化量，用来衡量编码计算复杂度，具体如式（5）所示，式中 $Time_{prop}(QP)$ 代表使用本文算法的实际编码时间，而 $Time_{VTM}(QP)$ 表示原始编码器 VTM10.0 的编码耗时，另外 4 种不同的量化参数（QP）分别为 22，27，32，37，本文将 4 种 QP 下编码时间缩减量取平均作为最终编码时间缩减量。

$$\Delta T = \frac{1}{4}\sum_{QP}\frac{Time_{VTM}(QP) - Time_{Prop}(QP)}{Time_{VTM}(QP)} \times 100\% \quad (5)$$

实验结果如表 4 所示，首先不执行基于 CNN 的上采样，预处理阶段采用的均是双三次插值上采样，此时观察到所提出的算法相较于原 VTM 在时间上大大减少，这证明了编码的计算复杂度大大的降低了，同时表中显示所有序列的编码性能相似，相较于原来平均可以节省 43.45%的时间，BD-BR 平均增加了 1.037%，表明该模型具有良好的泛化能力，提出的深度学习预测算法确实可以有效取代原本的筛选算法。从类别细分显示，该算法对于带有运动的序列或复杂情况下的混合内容序列的性能更好更明显，例如 MarketPlace、BasketballDrill、BQMall、RitualDance，

分别能带来 46.27%、44.74%、46.53%、45.39%的编码减少时间，带来这种结果的原因是基于这些类别中序列结构的复杂性，当视频序列过于复杂或者剧烈运动情况，正常的帧内预测模式会变得难以预测，而对于本文的算法来说，始终是分类问题。从序列的分辨率来看，高分辨序列相较于低分辨率在编码节省的时间上，效果也更加的明显，但同时付出的代价就是 BD-BR 的增加。

表 4 实验结果仿真表

分辨率	测试序列	没有 CNN 上采样		加入 CNN 上采样	
		BD-BR(%)	ΔT(%)	BD-BR(%)	ΔT(%)
A(3840×2160)	Campfire	1.328	-45.31	0.783	-40.73
	CatRobot	1.087	-43.89	0.717	-41.03
	FoodMarket	1.218	-46.57	0.828	-42.68
B(1920×1080)	ParkScene	1.477	-44.29	0.737	-37.85
	BQTerrace	1.124	-42.69	0.683	-36.45
	RitualDance	0.959	-45.39	0.692	-42.52
	MarketPlace	1.172	-46.27	0.734	-41.33
C(832×480)	BasketballDrill	1.015	-44.74	0.728	-43.04
	PartyScene	0.917	-40.64	0.471	-34.32
	RaceHorsesC	0.934	-42.47	0.604	-38.15
	BQMall	1.218	-46.53	0.653	-43.37
D(416×240)	BQSquare	1.017	-38.36	0.595	-34.47
	BlowingBubbles	0.907	-41.24	0.493	-36.27
	BasketballPass	0.772	-43.36	0.507	-40.17
	RaceHorses	1.017	-40.21	0.559	-37.56
E(1280×720)	Johnny	0.872	-45.29	0.616	-42.64
	KristenAndSara	0.953	-43.34	0.624	-40.08
	FourPeople	0.687	-41.58	0.581	-39.48
	平均性能	1.037	-43.45	0.645	-39.56

之后在预处理阶段加入本文设计的基于 CNN 的上采样模型，来提高原本的上采样图像质量，此时如表 4 所示，平均节省了 39.56%的编码时间，但 BD-BR 平均增加变为 0.645%，相较于不加入基于 CNN 的上采样模型，多花费了 3.89%的编码时间，减少了接近 0.40%的 BD-BR，从类别上来看，该上采样模型对于那些纹理清晰，纹理复杂的序列效果比较明显，例如 PartyScene、BlowingBubbles，它们的 BD-BR 分别变化为 0.471%、0.493%，相较于没有采用 CNN 上采样分别减少了 48%、45%，同时从分辨率来看，对于高分辨率的序列，影响更大，BD-BR 普遍都有明显的降低。因此以上结果说明基于 CNN 的上采样确实提高了算法编码的质量。

表 5 显示了本文的算法与最新的三种针对 VVC 的帧内快速模式算法相比的结果，各个算法它们都具有良好的 RD 性能，其中文献[13]和文献[12]是基于深度学习的方法，文献[13]通过 CNN 训练来预测每个 4*4 划分的概率，进而推导出划分模式；文献[12]根据 QTMT 灵活的结构，提出了具有早退机制的多阶段退出 CNN 模型。文献[5]是传统算法，选择纹理方向为基准作为快速帧内模式决策。 因为各个文献中测试序列的差异，因此取序列类结果的平均值进行比较。

现将本文算法与参考算法的性能比较总结如表 5 所示，相较于传统算法，本文的的算法不管在编码节省时间上还是 BD-BR 上，均有明显的优势；与其余深度学习的方法相比，虽然编码节省时间略有降低，但他们的 BD-BR 过高，本文在加入 CNN 上采样的时候能够保证较低的 BD-BR，所以总体性能上在所有算法中是最优的，说明本文算法是具有竞争力的。

表 5 本文算法与其他论文算法比较

序列类别	本文算法		文献[5]		文献[13]		文献[12]	
	BD-BR(%)	ΔT(%)	BD-BR(%)	ΔT(%)	BD-BR(%)	ΔT(%)	BD-BR(%)	ΔT(%)
A(3840×2160)	0.78	-41.48	0.58	-29.58	0.85	-53.84	1.47	-43.41
B(1920×1080)	0.71	-39.54	0.61	-28.57	0.75	-51.57	1.15	-45.92
C(832×480)	0.61	-39.72	0.48	-29.93	0.56	-26.45	1.18	-44.75
D(416×240)	0.54	-37.12	0.52	-31.38	0.33	-22.73	1.07	-42.53
E(1280×720)	0.61	-40.73	0.84	-30.31	1.18	-43.80	1.81	-49.27
平均性能	0.64	-39.56	0.61	-29.95	0.75	-39.68	1.34	-45.18

3 结论

在本文中，提出了一种针对 VVC 的基于深度学习的帧内模式决策快速算法。首先对于 H.266 的特点，建立了一个模式决策数据集，在预处理上通过插值滤波器和 CNN 上采样相结合的方式处理原本复杂多变的 CU 块，之后设计了一个与模式决策相适应的深度学习模型，除了当前编码块，引入了相邻参考块和残差块作为输入组件，最后将模型嵌入 VTM10.0 中，从而将帧内预测的模式决策转换化为分类问题，降低帧内模式预测复杂度。实验结果表明，本文的算法平均可减少 39.56%~43.45%的编码时间，BD-BR 仅增加 0.645%~1.037%。本文所提出算法是采用深度学习方法集成于 H.266/VVC 视频编码进行快速算法优化的有益尝试，相较于最新的几种传统快速算法也取得了较为明显的性能提升，在未来的研究中，可以将深度学习的方法拓展到 VVC 的帧间预测、环路滤波等模块，从而进一步的提高编码效率或降低复杂度。

参考文献

[1] Filippov A, Rufitskiy V. Recent Advances in Intra Prediction for the Emerging H. 266/VVC Video Coding Standard[C]//2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON). IEEE, 2019: 0525-0530.

[2] Bross B, Chen J, Liu S, et al. Versatile video coding (draft 10)[J]. ITU-T and ISO/IEC JVET-S2001, 2020.

[3] Siqueira Í, Correa G, Grellert M. Rate-distortion and complexity comparison of HEVC and VVC video encoders[C]//2020 IEEE 11th Latin American Symposium on Circuits & Systems (LASCAS). IEEE, 2020: 1-4.

[4] Yang H, Shen L, Dong X, et al. Low-complexity CTU partition structure decision and fast intra mode decision for versatile video coding[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 30(6): 1668-1682.

[5] Zhang Q, Wang Y, Huang L, et al. Fast CU partition and intra mode decision method for H. 266/VVC[J]. IEEE Access, 2020, 8: 117539-117550.

[6] Zhao Z, Wang S, Wang S, et al. Enhanced bi-prediction with convolutional neural network for high-efficiency video coding[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 29(11): 3291-3301.

[7] Lee J K, Kim N, Cho S, et al. Deep video prediction network-based inter-frame coding in HEVC[J]. IEEE Access, 2020, 8: 95906-95917.

[8] Ma C, Liu D, Peng X, et al. Convolutional neural network-based arithmetic coding for HEVC intra-predicted residues[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 30(7): 1901-1916.

[9] Wang M Z, Wan S, Gong H, et al. Attention-based dual-scale CNN in-loop filter for Versatile Video Coding[J]. IEEE Access, 2019, 7: 145214-145226.

[10] Li J, Li B, Xu J, et al. Fully connected network-based intra prediction for image coding[J]. IEEE Transactions on Image Processing, 2018, 27(7): 3236-3247.

[11] Hu Y, Yang W, Li M, et al. Progressive spatial recurrent neural network for intra prediction[J]. IEEE

Transactions on Multimedia, 2019, 21(12): 3024-3037.
[12] Li T, Xu M, Tang R, et al. DeepQTMT: A Deep Learning Approach for Fast QTMT-based CU Partition of Intra-mode VVC[J]. IEEE Transactions on Image Processing, 2021, 30: 5377-5390.
[13] Tissier A, Hamidouche W, Vanney J, et al. CNN oriented complexity reduction of VVC intra encoder[C]//2020 IEEE International Conference on Image Processing (ICIP). IEEE, 2020: 3139-3143.
[14] Lin T L, Liang K W, Huang J Y, et al. Intra mode prediction for H. 266/FVC video coding based on convolutional neural network[J]. Journal of Visual Communication and Image Representation, 2020, 71: 102686.
[15] Wang Z. Fast depth coding in 3D-HEVC using deep learning[D].Xiang Gang:The Hong Kong Polytechnic University,2018.
[16] 李萍,刘以安,徐安林. 基于多尺度耦合的密集残差网络红外图像增强[J].电子测量与仪器学报,2021, v.35;No.247(7):148-155.
[17] Jiang K, Wang Z, Yi P, et al. Multi-scale progressive fusion network for single image deraining[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 8346-8355.
[18] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.
[19] 杨梅,贾旭,殷浩东,等.基于联合注意力孪生网络目标跟踪算法[J].仪器仪表学报,2021,42(1): 127-136.
[20] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(9): 1904-1916.
[21] Bossen F. Common test conditions and software reference configurations[J]. JCTVC-L1100, 2013, 12(7).
[22] 郭 泽 .H.266/VVC 高效帧间预测编码算法研究[D]. 西安：电子科技大学,2019.

作者简介

施金诚（通讯作者），工学硕士，主要研究方向为视频编解码。
E-mail：shijinchhen@163.com
杨静，工学博士，副教授，主要研究方向为图像通信和视频编解码。
E-mail：jingyang@shmtu.edu.cn