DOI:10.19651/j.cnki.emt.2211463

基于卷积神经网络的分数像素运动补偿*

郑乐佳[1] 郝禄国[1] 项 颖[1] 曾文彬[2]
(1.广东工业大学信息工程学院 广州 510006;2.天津大学电气自动化与信息工程学院 天津 300192)

摘 要: 分数阶插值是帧间编码运动补偿中一项重要技术。为改进传统插值滤波器插值效果不佳,现有基于深度学习的方法存在只生成半像素样本、需要对各个分像素位置及量化参数(QP)训练相应模型、需引入额外的信息作为输入等不足之处,本文提出一种用于帧间编码分数像素运动补偿的卷积神经网络(CNN)方法。首先以残差稠密网络为基础,然后联合多尺度失真特征提取结构及亚像素卷积来增加特征提取准确性和生成分数像素。为了训练所提出的网络,本文分析该分数阶插值任务的特点,构建了带有真实性失真的数据集。该模型依靠参考帧生成各个位置的分数像素样本,且适应任意的量化参数。实验结果表明,与 H.265/HEVC 相比可以节省更多的比特数。在低延迟 P(LDP)的配置下,平均降低 2%的 BD-rate,与同类方法相比综合性能也有所提升。

关键词: H.265/HEVC;帧间预测;分数像素运动补偿;CNN
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

Convolutional neural network-based fractional-pixel motion compenstaion

Zheng Lejia[1] Hao Luguo[1] Xiang Ying[1] Zeng Wenbin[2]
(1. School of Information Engineering, Guangdong University of Technology, Guangzhou 510006, China;
2. School of Electrical Automation and Information Engineering, Tianjin University, Tianjin 300192, China)

Abstract: A convolutional neural network (CNN) for fractional interpolation of inter prediction is proposed because of the poor interpolation effect of traditional interpolation filters and the deep learning methods, which only generate half pixel samples, or need to train the corresponding model for each pixel position and quantization parameter (QP), or introduce additional information as input. Based on the dense residual network, the model combines multi-scale distortion feature extraction structure and sub-pixel convolution to increase the accuracy of feature extraction and generate fractional pixels. The characteristics of fractional interpolation task are analyzed and the data set with true distortion is constructed. The model directly generates fractional pixel samples and can adapt to arbitrary quantization parameters (QP). Experimental results verify the efficiency of the method. Compared with H.265/HEVC, this method achieves 2% in bit saving on average under low-delay P configuration. Compared with similar methods, the overall performance has also been improved.

Keywords: H.265/HEVC;inter prediction;fractional-pixel motion compensation;CNN

0 引 言

基于块的运动补偿是目前主流的视频编码框架的关键技术之一。现有的视频编码标准,包括 H.265/HEVC[1]均采用此方法进行帧间预测编码,以减少时间冗余,显著提高了编码效率。在运动补偿中,对于当前要编码的块,在之前已编码后再重构的参考帧中寻找最佳匹配的块作为预测块,然后将两块之间的残差和运动向量(motion vector, MV)进行编码。

由于数字视频在空间上的离散采样,意味着整数位置的参考块可能与当前块不能很好地匹配。在预测当前块时,仅对参考帧进行搜索是不够的。为了寻找更适合的参考块,通过对检索到的整数位置样本进行分数阶插值,在分数像素位置生成参考样本,然后搜索最佳匹配块作为当前块的预测块。因此分数像素运动补偿方法被广泛用于视频编码标准中,如 H.264/AVC[2]和 HEVC 都使用到了 1/4 像素精度的分数像素运动补偿。

传统的分数像素插值采用固定插值滤波器,HEVC 采

收稿日期:2022-09-20
* 基金项目:广东省自然科学基金(2022A1515010777)、广东省科技计划项目(2022A0505050072)资助

用基于离散余弦变换的插值滤波器(discrete cosine transform, DCTIF),对亮度分量的 1/4 精度插值采用 7 抽头系数,对 1/2 精度插值采用 8 抽头系数[1]。这样的滤波器通常是通过信号处理理论推导出来,假设要插值的信号带限。尽管计算简单,这种滤波器并不能很好地处理不同种类内容的视频,因为自然视频中的内容比理想的带限信号复杂得多。同时,这种滤波器只依靠相邻的几个像素点进行插值,插值结果的质量有限。因此固定插值滤波器不能很好适应各种不同结构和内容的视频信号。

近年来,随着深度学习的不断发展,许多基于卷积神经网络的方法被提出用于处理计算机视觉任务,如图像分类[3],去噪[5]和目标检测[6]。同时,基于深度学习的超分辨率方法也逐渐出现。文献[7]首先提出了一种单图像超分辨率神经网络(SRCNN),该方法直接学习低分辨率到高分辨率图像之间的端到端映射。随后,文献[8]提出了一个更深的带有残差学习的神经网络(VDSR),进一步提高了超分辨率性能。这些深度学习方法在图像处理和计算机视觉任务中的成功启发了基于深度学习的视频压缩方法。

目前基于深度学习的视频编码方法主要分为两大类:其一是端到端的编码模型,即使用网络模型作为整个编码模型替代传统编码框架;其二是沿用传统编码框架,将特定网络模型集成到框架上以提高效率,这也是本文所讨论的方法。此类方法有降低复杂性[9]、去除压缩伪影[10],视频转码[11]和帧内预测[12]等。由于受到图像超分辨率任务的启发,研究者也将深度学习方法应用于视频帧间预测编码的相关工作中。

文献[13]首先提出了基于卷积神经网络的插值滤波器来代替 HEVC 的 1/2 像素插值滤波器。随后,文献[14]将超分网络 VDSR 用于 HEVC 对 1/2 像素插值滤波器性能进行了改进。虽然可以取得不错的效果,但是这两种方法只考虑了 1/2 像素精度,并没有针对 1/4 精度的插值滤波器,还需为每个 1/2 像素位置训练一个特定的网络。这不仅增加了复杂度还不利于神经网络在编码上的应用。因此,文献[15]提出一种群变分变换卷积神经网络(GVTCNN),用一个网络来推断不同分数像素位置的样本。同样的,文献[16]提出了一种基于切换模式的深度分数插值方法,从不同位置的整数像素推断出子像素,对 GVTCNN 进行了改进,减少了运动移位的缺点。而文献[17]提出一种基于更深层次分组变异的网络(GVCNN),该方法在处理不同量化参数(QP)和分数像素位置时具有通用性。与之前都不同的是,文献[18]将分数像素运动补偿问题表示为图像间的回归问题,并提出采用 CNN 模型来处理该问题。文献[19]使用参考块上的实整数位置的样本来预测和生成更加接近当前编码块的分数像素样本。文献[20]复用超分网络 FSRCNN,将参考帧与相邻分量的残差作为输入,较好的实现不同 QP 下的分数插值。这些基于 CNN 的方法都取得了不错的效果。本文的工作也着重于利用 CNN 提高分数像素插值的精度。

现有的基于深度学习的方法虽然取得了良好的效果,但是多数方法要么需要给每个分数位置训练模型,要么需要对不同 QP 训练对应的模型,要么需要引进额外的辅助信息。为此,本文构建了一个用于分数像素运动补偿的神经网络模型,使用参考帧作为输入,用于生成 1/2 精度和 1/4 精度的像素样本,能够适应不同 QP 下的插值需求,经实验证明该方法可以有效的节省比特数提高编码质量。本文的模型构建步骤包括:

1)为了去除参考帧经过压缩重建带来的压缩伪影与量化噪声,在模型开头部分采用变尺寸滤波器技术[21]来捕获失真信息的多尺度相似性,以减少各类噪声的影响,使主干网络能更好的进行深度特征提取。

2)虽然与超分辨率任务相似,但是分数插值只需要生成分数样本,对于原来的整数位置不需要改变[14]。本文在模型末尾上采样部分没有采用反卷积层(deconvolution layer)而是采用亚像素卷积(sub-pixel convolution)[26],经分析认为这会更加符合分数像素插值任务的特性。

3)此外,还对模型参数进行调整,并进行对比测试。同时,由于分数样本不是真实存在的[13],本文没有单纯使用现有的超分数据集,而是构建更加符合实际帧间编码分数阶插值任务需求的数据集。

4)在训练时,由于亚像素卷积的特点,对于 1/2 及 1/4 精度的样本需要不同的缩放倍数,因此需要训练缩放倍数为 2 和 4 的两个网络对应两种精度的插值,然后再将训练好的模型集成到 HEVC 中,即添加到 HM 程序的帧间预测编码的运动补偿部分,测试其视频编码性能。实验结果表明了该方法的有效性和优越性。

本文的其余部分如下:在第 1 节描述了所提方法的细节,第 2 节介绍了训练数据的生成过程及网络的训练配置,第 3 节给出了测试细节和实验结果,第 4 节是本文总结。

1 基于 CNN 的分数像素插值方法

1.1 方法概述

本文提出了一种用于运动补偿的基于 CNN 的分数像素插值方法。整像素和分数像素的位置如图 1 所示。其中,$I_{i,j}$ 代表的是整数像素位置,而 $h^k_{i,j}\{k=1,2,3\}$ 是 1/2 像素位置,$q^k_{i,j}\{k=1,2,\cdots,12\}$ 代表 1/4 像素位置。在编码过程中,将给定的参考帧做为整数样本 $I_{i,j}$,然后通过插值滤波器,从整数样本插值出 1/2 像素位置和 1/4 像素位置。最终在插值后的样本搜索最接近当前待编码帧的部分进行编码。传统方法,如 HEVC 采用 DCTIF 进行插值。在本文中,训练了两个网络模型分别对应 1/2 及 1/4 精度的插值任务。

虽然分数像素插值任务与图像超分辨率相关,但还是存在不同之处:其一,插值任务目的是生成分数样本,而整数位置的像素保持不变;其二,虽然都是从低分辨率生成高

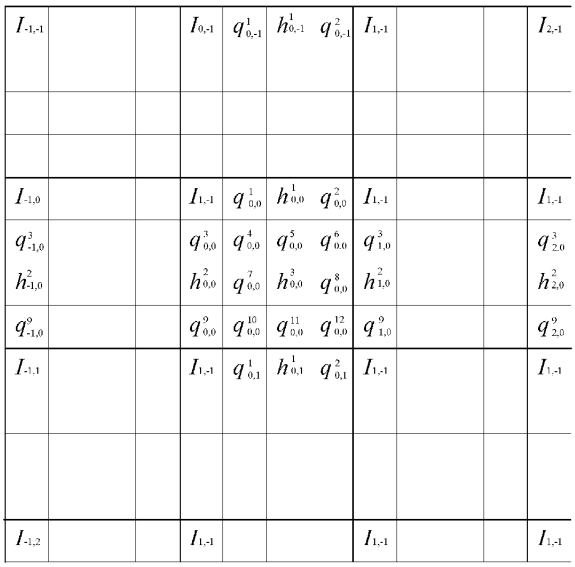

图 1　不同分数像素位置

分辨率，但待插值的参考帧是经过编码重建的，带有压缩伪影和量化噪声，并且当量化参数 QP 越大，噪声越大。因此，不能简单的复用现有的图像超分辨率的方法，而且有关实验[13]也证明了直接使用超分辨率的方法不能很好发挥作用。

1.2　网络模型

基于现有的图像超分辨率方法和分像素插值任务的需求，本文构建了相应的网络模型如图 2 所示。该模型主要可以分为三部分：多尺度失真特征提取网络，主干特征提取网络和上采样网络。

每层的输出都是通过前一层的线性变换得到，激活函数采用的是 LeakyReLU（LReLU）[22]，其公式如下：

$$\mathrm{LeakyReLU}(x)=\begin{cases}x, & x\geqslant 0\\ \alpha x, & \text{其他}\end{cases} \tag{1}$$

LReLU 作为 ReLU 的变体，对于输入小于 0 部分不再直接等于 0，而是可以设置 α 的值保证在输入小于 0 的时候有微弱的输出，减轻了 ReLU 的稀疏性和神经元死亡的问题，更适合于图像生成任务。

1）多尺度失真特征提取网络

在图 2 所示的网络结构中，图像输入的第一部分是一个多尺度的失真特征提取结构，用于提取浅层特征和消除压缩伪影及噪声。采用的是经过修改的 Inception 模块[23]，将 Inception 模块中的批量标准化（BN）移除，使结构成为一个全卷积结构。由于 BN 层在训练期间使用批次的均值和方差对特征进行归一化，在测试期间使用整个训练数据集的估计均值和方差，当训练和测试数据集的统计数据差异很大时，BN 层往往可能会引入不适的伪影，限制了泛化能力。

图像输入后是经过 3 个不同尺寸的卷积层，所采用的卷积核尺寸大小分别 1×1、3×3 和 5×5。这些不同大小的卷积核意味着感受野不同，特征的映射区域大小也不同。通过拼接将不同尺寸特征进行合并，可以减少噪声的干扰。

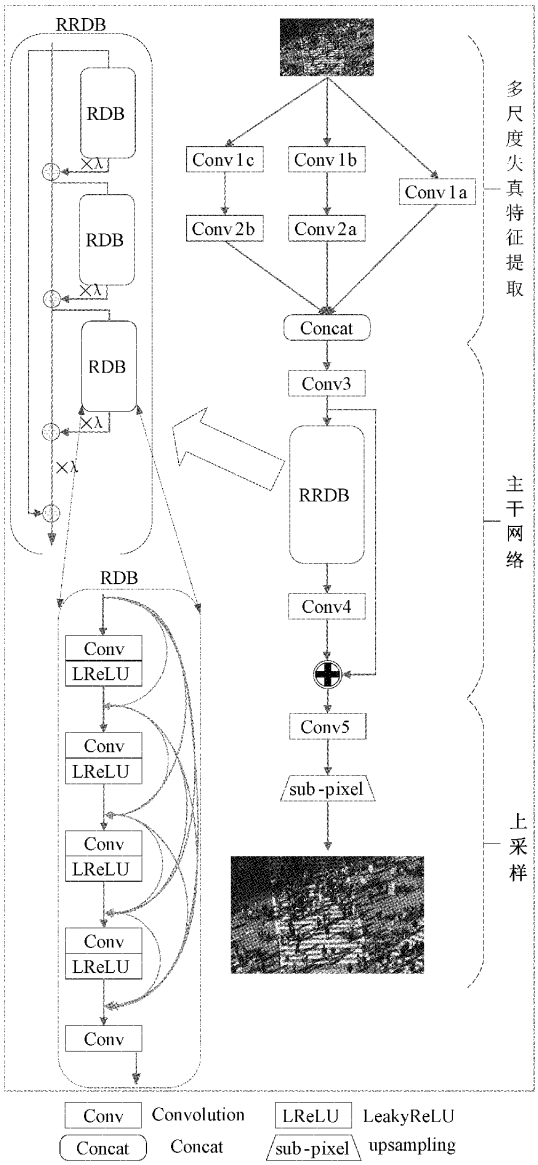

图 2　网络结构图

由于参考帧是通过编码量化后重建的，其存在的压缩伪影和量化损失所产生的噪声会影响网络的特征提取，而且量化参数 QP 不同，不同参考帧的损失也不同。如果不消除失真，直接提取特征输入到重构层，插值后的图像会包含许多不必要的压缩伪影和噪声，从而限制了插值的性能。HEVC 编码单元（CU）的划分为 8×8～64×64 的多个尺度，重建帧所带的失真是多尺度的。采用多尺度的输入特征提取能够更好的去除噪声影响，以使得主干特征提取网络不受噪声影响。

2）主干特征提取网络

在经过多尺度的特征提取合并后，会先通过一个 3×3 的卷积层对特征进行融合，用于主干特征提取网络的输入。

主干网络部分提取的特征会用于最后上采样产生插值结果,因此能否较好的提取参考帧的像素特征及关联性会直接决定插值结果的准确性。

由于残差稠密网络[24]在图像超分取得良好的结果,本文在该部分采用了残差密集(RRDB)模块[25]。然后再提取经过该模块的残差值,即 RRDB 模块所提取的特征作为前面的融合特征的残差进行合并。相比较于传统的稠密模块,该网络结构也去掉 BN 层,原因与上小节所阐述一致,这有利于去除伪影,提高泛化能力。

从图 2 可以观察到,RRDB 采用了两层残差结构,该结构由一个大的残差结构构成,核心部分由 3 个残差稠密块(RDB)构成,将核心网络的输出与残差边叠加,通过稠密连接卷积层提取丰富的局部特征,从先前 RDB 的状态直接连接到当前 RDB 的所有层,然后利用 RDB 的局部特征融合自适应地从先前和当前的局部特征中学习更有效的特征。RRDB 模块可以表示为如下:

$$\begin{aligned} I_1 &= F(I_{in}) \\ I_2 &= F(I_1) \\ I_3 &= F(I_2) \\ I_{out} &= \lambda \cdot (I_1 + I_2 + I_3) + I_{in} \end{aligned} \tag{2}$$

式中:I_{in} 代表输入特征,I_{out} 代表输出特征,F 代表残差稠密块 RDB,λ 代表残差缩放系数。将残差的分支乘上系数 λ,再加到主干上,可以使训练更加稳定。

在该模块中,卷积层从前面的所有层获取额外的输入,并将自己的特征映射传递给后面的所有层,该结构适合于在视频编码中捕捉多尺度特征。

在多特征叠加之后,使用一个 3×3 的卷积层作为一个非线性特征映射层,目的是将提取的低分辨率图像的特征转换为高分辨率图像的特征。

3) 上采样网络

网络的最后部分用于上采样,与一般方法不同的是,本文在分析了分数像素插值任务后认为反卷积操作会存在大量补 0 的区域,这对插值任务的结果可能是有害的,而使用亚像素卷[26]更适合该任务的上采样部分。亚像素卷积的示意图如图 3 所示。

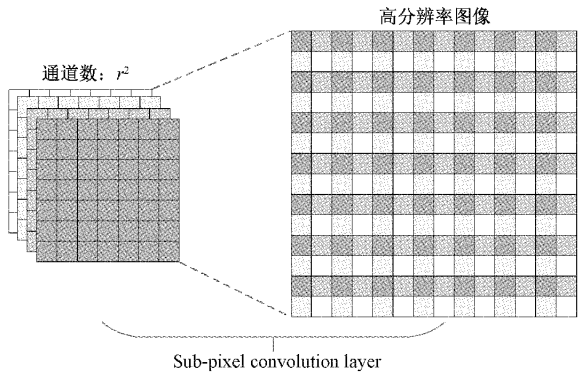

图 3 亚像素卷积操作

亚像素卷积输入的特征通道数为 r^2,其中 r 为缩放倍数,该部分可以表示为:

$$I^{SR} = SP(W_L \cdot f^{L-1}(I^{LR}) + b_L) \tag{3}$$

式中:SP 代表亚像素卷积操作,W 和 b 分别为第 L 层的参数,f 为 $L-1$ 层的卷积操作,I^{LR} 和 I^{SR} 分别表示低分辨率和高分辨率图像。该操作将宽高分别为 w 和 h,通道数为 Cr^2 的特征转换成宽为 rw,高为 rh,通道数为 C 的高分辨率图像。

由于亚像素卷积操作的实现过程不是直接通过线性插值等方式产生高分辨率图像,而是通过卷积先得到 Cr^2 个通道的特征图(特征图大小和输入低分辨率图像一致),然后通过周期筛选的方法得到这个高分辨率的图像,其中的 r 为上采样因子,也就是图像的扩大倍率,该方法能够有效的利用前面特征提取主干网络所提取到的特征来生成更高分辨率的图像,更适合本文任务的上采样部分。

4) 网络模型的配置

如表 1 所示为网络的具体配置,其中 RRDB 模块由 3 个 RDB 模块级联组成,每个 RDB 块包含 5 个 3×3 的卷积层,前四个卷积后会经过 LReLU 激活函数。每个 RDB 没有简单的直接相加,而是乘以残差因子 λ 再相加,这样不仅使模型训练更加简单也使得模型更加稳定。

在实验过程中,通过不断调整参数来训练模型,目的是最小化训练集上的损失函数。在模型训练过程中,采用的是绝对值误差即 L1 损失函数。设置 H 为生成分数像素的神经网络,θ 代表卷积核、偏置等待学习的参数。损失函数可以表示为:

$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n} \| H(x_i;\theta) - y_i \| \tag{4}$$

其中,x_i 和 y_i 分别代表低分辨率和高分辨率的图像对。

表 1 网络结构的配置

结构		卷积核	通道数
Layer1	Conv1a	1×1	64
	Conv1b	1×1	64
	Conv1c	1×1	48
Layer2	Conv2a	3×3	96
	Conv2b	5×5	64
Layer3	Conv3	3×3	64
Layer4	Conv4	3×3	64
Layer5	Conv5	3×3	64
RRDB	Conv	3×3	32
	LReLU	—	—

2 数据集生成及训练配置

2.1 数据集生成

自建的模型训练集和验证集所用到的图片数据集及视

频序列如表 2 所示。由于视频编码中的分数阶插值所处理的参考帧带有失真，因此生成具有真实失真的样本来进行模型训练是非常重要的。本文任务的分数阶插值方法的有效性需要通过最终的编码性能来定量评价，在复杂的视频编码系统中，编码性能受很多因素的影响。这意味参考帧所带损失也受多因素影响，而且插值的输入没有确切的高分辨率图像作为模型输出结果的真实标签。

表 2　数据集的配置

训练集		验证集	
图像集	T91	图像集	Set5
视频集	football	视频集	silent
	deadline		students
	bowing		shields
	park joy		—
	rushhour		—
	pedestrian		—
	parkrun		—
	stockholm		—
	crowd run		—

在之前的方法[13]中，许多方法均采用单张的图像进行高斯模糊后下采样得到整数像素样本，然后将该子图像经过 HEVC 编码重建得到训练集的输入，而将原图作为结果标签，以此来模拟真实情况下的输入数据。但对单张图像只能使用帧内编码，帧内编码和帧间编码采用不同的编码方式，压缩单张图像中的压缩伪影不能完全模拟帧间编码的特征。而且在本文任务中的基于 CNN 的插值滤波器应该要进一步消除失真。由于经过帧内编码的压缩图像不能模拟帧间编码的失真特性，如果简单地在这些图像数据集上训练网络，模型并不能很好的适应分数阶插值任务，且会引入与任务无关的误差。因此，应该从经过帧间编码后的视频序列中抽取帧作为数据集，将对应的未编码的序列的同一帧作为结果的标签。但单一序列的帧之间具有很强的相似性，这样生成的数据集容易过拟合且模型泛化能力不足，需要从多个不同类型的序列抽取帧。

在自建的数据集中，本文从多个种类视频序列图像中提取帧以提供各种细节，并具有足够多的重复增强特征。数据集使用了 8 个原始序列用于训练数据集的生成，3 个序列用于验证数据集的生成。所采用的序列分辨率大小范围从 352×288 到 1 920×1 080。对于每个序列，需要在多个量化参数下编码重建，然后每 10 帧提取一帧作为数据以减少相似性。并且，在训练模型前，使用现有的超分数据集 T91 作为训练集，数据集 Set5 作为验证集来对模型做预训练，使模型有一个好的特征提取能力和泛化能力。在超分数据集预训练的基础上，再采用自建的数据集做进一步训练，以使得模型能够满足分数像素插值任务要求。由于上采样的倍数不同，使得模型在上采样层有不同的特征通道，因此需要对 1/2 和 1/4 精度各训练一个模型。由于只在 Y 分量上进行训练，模型用于 Y 分量的插值任务，可以应对任意的 QP 要求。

2.2　训练配置

本文采用 pytorch 框架来训练模型。数据集中输入的图片被裁为 20×20 的大小对应 1/2 精度插值模型的输入，16×16 大小对应 1/4 精度。对应的标签图像相应为 40×40 和 64×64。由于数据量足够多，因此不需要进行图像增强，最终得到的训练数据集，1/2 精度插值模型有 690 688 个样本，1/4 精度插值模型有 258 400 个样本。模型训练时采用 Adam 优化器，学习率设置为 0.000 1。RRDB 模块中残差缩放因子 λ 和 LReLU 的 α 值参考文献[25]的实验设置。训练模型时所采用的服务器 CPU 配置为 Intel Xeon Silver 4210R，GPU 为 RTX A5000。

3　测试细节和实验结果

3.1　测试过程

在 HEVC 中，预测单元 PU 是帧间编码的基本单元，当对预测单元进行运动估计时，重用现有的运动估计算法，在进行分像素插值时才调用基于 CNN 的模型。如图 4 所示为本文算法集成到 HEVC 框架的示意图，需要将训练好的模型集成到 HEVC 参考软件 HM(版本 16.15)中。在使用过程中，本文没有直接替换原有的分像素运动补偿方法，而是在编码单元 CU 进行率失真选择，判断是使用网络模型还是原有的 DCTIF，对同一个 CU 里的 PU 则会采用相同的插值方法。在编码器端选择不同的分数阶参考方法，然后将所选分数参考方法的标志位也进行编码，以供解码器端解码时选择对应的方法。

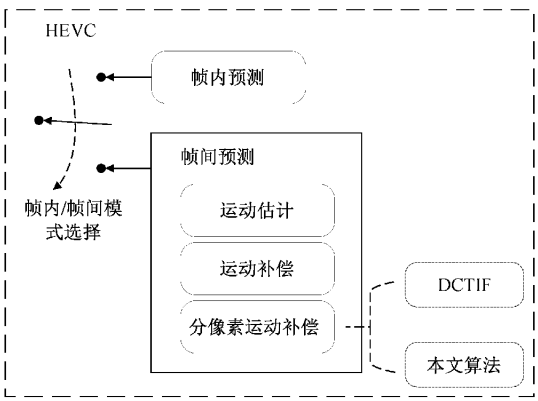

图 4　本文算法集成至 HEVC 的位置

HM 程序下采用 LDP 配置在 QP 为 22，27，32，37 四种量化参数下对 HEVC 官方提供的通测序列进行编码测试。在评价本文的方法时，码率和峰值信噪比(PSNR)都需要考虑，因此采用 BD-rate 来作为率失真的评价指标，BD-rate 值为负则表示优化后算法的编码性能得到了提高。

3.2 实验结果

1) 本文模型

在 HEVC 提供的通测序列类别为 B、C、D、E 和 F 下我们所提方法与 HEVC 标准方法的比较结果如表 3 所示。由表格可以看出,在 LDP 配置下,平均可以降低 2%的 BD-rate,获得不小的提升。其中,序列 BQTerrace 和 BasketballPass 降的最多,达到了 3%。其原因是这两个序列均为非均质、运动较剧烈的视频序列。BQSquare 序列由于其结构均质且平坦,只降低了 0.8%。这说明本文所提方法在纹理复杂的序列上可以获得更好的效果,其原因是模型能够生成更加接近待编码图像的预测块,而传统的方法依靠固定插值滤波器会造成图像模糊,因此在纹理复杂序列上的效果更差。

表 3 LDP 配置下本文算法的 BD-rate

类别	测试序列	3 个分量的 BD-rate		
		Y/%	U/%	V/%
B	kimono	−1.4	0.4	0.5
	BQTerrace	−3.0	−0.7	−1.0
	BasketballDrive	−1.3	−0.7	−0.6
	ParkScene	−1.3	−0.7	−0.6
	Cactus	−1.5	−0.8	−0.7
	Average	−1.7	−0.5	−0.5
C	BasketballDrill	−2.9	−1.2	−0.8
	BQMall	−2.8	−1.3	−1.4
	ParkScene	−1.8	−0.9	−1.1
	RaceHorsesC	−1.9	−1.1	−1.2
	Average	−2.3	−1.1	−1.1
D	BasketballPass	−3.0	−2.2	−1.5
	BlowingBubbles	−2.3	−0.8	−0.4
	BQSquare	−0.8	0.8	0.6
	Racehorses	−2.1	−1.2	−1.1
	Average	2.0	−0.8	−0.6
E	FourPeople	−1.9	−0.1	−0.5
	Johnny	−2.3	−0	−0.7
	KristenAndSara	−2.4	−0.6	−0.4
	Average	−2.2	−0.2	−0.5
F	BasketballDrillText	−1.9	−1.0	−0.8
	ChinaSpeed	−1.7	−2.0	−1.9
	SlideEditing	−1.7	−1.2	−1.0
	SliceShow	−1.9	−0.9	−0.7
	Average	−1.8	−1.2	−1.1
ALL	Overall	−2.0	−0.8	−0.8

为了直观的比对,在图 5 给出了 4 条 R-D 曲线,其横坐标为码率,纵坐标为 PSNR,用来对比衡量方法好坏。涵盖了 4 个不同分辨率、不同类别的视频序列,包括 B 类的 ParkScene 序列,C 类的 BasketballDrill 序列,D 类的 BlowingBubbles 序列以及 F 类的 ChinaSpeed 序列。图中实线代表本文的方法,虚线代表 HEVC。明显可以看出本文方法确实比原来方法较好,但在不同 QP 下,模型所带来的效益是不同的。在高码率时比在低码率时模型性能更好,这可能是参考帧图片在较低的码率下包含更多的压缩噪声,而这些压缩噪声是高度非线性的,使得模型特征提取的效果受到影响,导致最终生成的预测帧质量较差。

2) 对比试验

为了对所提方法性能能够有更直观的比较,将本文的方法与现有的基于深度学习的分数插值方法,文献[17]、[20]、[27]进行了比较。为了更准确的比较,还将本文的方法集成到 HM 16.4 中。

在这个比较中,所用到的是复用文献[20]中复现的结果。3 种方法均在 HEVC 通用测试条件下进行测试,结果都为与 HEVC 标准的对比,均采用 Y 分量的 BD-rate 作为率失真的评价指标,同在 LDP 配置下,哪个方法的 BD-rate 值越小,证明所用方法对编码性能提高得越明显。如表 4 所示可以看出,本文的方法在多数情况下都比其他两种方法获得了更好的增益。

表 4 LDP 配置下本文算法与其他算法 BD-rate 对比

%

类别	测试序列	文献[17]	文献[20]	文献[27]	本文方法
C	BasketballDrill	−2.2	−2.4	−1.8	**−2.9**
	BQMall	−1.7	−2.8	−0.6	**−2.8**
	ParkScene	−1.5	−1.7	0.2	**−1.8**
	RaceHorsesC	−1.9	−1.9	−0.8	**−1.9**
	Average	−2.1	−2.2	−0.9	**−2.3**
D	BasketballPass	−2.4	**−3.2**	−2.4	−3.0
	BlowingBubbles	−1.9	−2.1	−0.4	**−2.3**
	BQSquare	−0.6	−0.8	0.5	**−0.8**
	Racehorses	**−2.6**	−2.5	−1.0	−2.1
	Average	−1.9	−1.9	−0.8	**−2.0**
E	FourPeople	−1.4	−1.7	−1.3	**−1.9**
	Johnny	−2.5	**−2.9**	−1.4	−2.3
	KristenAndSara	−2.2	−2.1	−1.0	**−2.4**
	Average	−2.0	−2.2	−1.2	**−2.2**

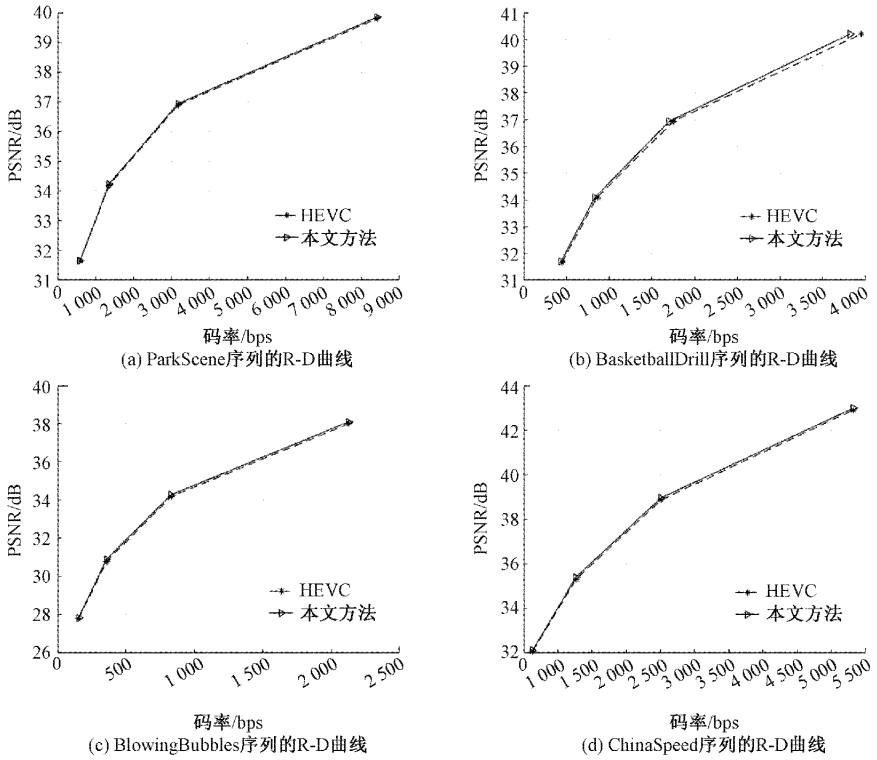

图 5　本文算法与 HEVC 的 R-D 曲线对比

4　结　　论

生成更准确的参考帧对帧间预测编码效率的提高是至关重要的一步。本文提出了一种基于卷积神经网络的分数像素运动补偿方法，通过生成更加准确的分数像素来提高帧间预测效率。对于所输入的参考帧，先经过多尺度特征提取结构来提取浅层特征和消除失真，再通过多级稠密残差网络提取所需的深层特征，最后使用亚像素卷积进行上采样生成所需预测帧。相关实验结果证明，与 HEVC 相比该方法显著节省了码流，验证了该网络在分数像素运动补偿任务上的有效性。但由于该方法只使用单一的参考帧，没有考虑双向的分数像素运动补偿，帧间预测的精度仍有进一步提高的空间。同时，只在 Y 分量训练了模型，还没有完全利用图像信息。在未来的工作中将会尝试将所提的方法扩展到双向运动补偿，并对颜色分量 U 和 V 也训练相应的模型，以获得更好的压缩效率和预测精度。

参考文献

[1] SULLIVAN G J, OHM J R, HAN W J, et al. Overview of the high efficiency video coding(HEVC) standard[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2012, 22(12): 1649-1668.

[2] VETRO A, WIEGAND T, SULLIVAN G J. Overview of the stereo and multiview video coding extensions of the H.264/MPEG-4 AVC standard[J]. Proceedings of the IEEE, 2011, 99(4): 626-642.

[3] MAHMOOD A, BENNAMOUN M, AN S, et al. ResFeats: Residual network based features for underwater image classification[J]. Image and Vision Computing, 2020, 93: 103811.

[4] XUE Z X, YU X C, BING L, et al. Hresnetam: Hierarchical residual network with attention mechanism for hyperspectral image classification[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 3566-3580.

[5] THAKUR R S, YADAV R N, GUPTA L. State-of-art analysis of image denoising methods using convolutional neural networks [J]. IET Image Processing, 2019, 13(13): 2367-2380.

[6] ROY A M, BOSE R, BHADURI J. A fast accurate fine-grain object detection model based on YOLOv4 deep neural network [J]. Neural Computing and Applications, 2022, 34(5): 3895-3921.

[7] CHAO D, CHEN C L, HE K, et al. Learning a deep convolutional network for image super-resolution[C]. European Conference on Computer Vision(ECCV), 2014: 184-199.

[8] KIM J, LEE J K, LEE K M. Accurate image super-resolution using very deep convolutional networks[C]. IEEE Conference on Computer Vision and Pattern

Recognition, 2016: 1646-4654.

[9] XU M, LI T Y, WANG Z L, et al. Reducing complexity of hevc: A deep learning approach[J]. IEEE Transactions on Image Processing, 2017, 27(10): 5044-5059.

[10] YANG R, XU M, WANG Z L, et al. Multi-frame quality enhancement for compressed video[C]. IEEE International Conference on Computer Vision and Pattern Recognition(CVPR), 2018: 6664-6673.

[11] XU J, XU M, WEI Y, et al. Fast h.264 to hevc transcoding: A deep learning metho [J]. IEEE Transactions on Multimedia, 2019, 21 (7): 1633-1645,.

[12] 施金诚,杨静. 基于深度学习的 VVC 快速帧内模式预测[J]. 电子测量技术,2022,45(3):104-111.

[13] YAN N, LIU D, LI H Q, et al. A convolutional neural network approach for half-pel interpolation in video coding[C]. 2017 IEEE International Symposium on Circuits and Systems(ISCAS), 2017: 1-4.

[14] HAN Z, LI S, LUO Z Y, et al. Learning a convolutional neural network for fractional interpolation in HEVC inter coding[C]. 2017 IEEE Visual Communications and Image Processing(VCIP), 2017: 1-4.

[15] XIA S, YANG W H, HU Y Y, et al. A group variational transformation neural network for fractional interpolation of video coding[C]. 2018 Data Compression Conference, 2018: 127-136.

[16] XIA S F, YANG W H, HU Y Y, et al. Switch mode based deep fractional interpolation in video coding[C]. 2019 IEEE International Symposium on Circuits and Systems(ISCAS), 2019: 1-5.

[17] LIU J Y, XIA S F, YANG W H, et al. One-for-All: grouped variation network-based fractional interpolation in video coding[J]. IEEE Transactions on Image Processing, 2019, 28(5): 2140-2151.

[18] YAN N, LIU D, LI H, et al. Convolutional neural network-based fractional-pixel motion compensation[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(3): 840-853.

[19] ZHANG H, LI L, SONG L, et al. Advanced CNN based motion compensation fractional interpolation[C]. 2019 IEEE International Conference on Image Processing (ICIP), 2019: 709-713.

[20] CHEN Z S, LIU J R, YANG J, et al. Super-resolution network-based fractional-pixel motion compensation[J]. Signal Image and Video Processing, 2021, 15: 1547-1554.

[21] DAI Y Y, DONG L, FENG W. A convolutional neural network approach for post-processing in hevc intra coding [C]. International Conference on Multimedia Modeling, 2017: 28-39.

[22] PARHI R, NOWAK R D. The role of neural network activation functions [J]. IEEE Signal Processing Letters, 2020, 27: 1779-1783,.

[23] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2016: 2818-2826.

[24] LIU C C, SUN X F, CHEN C Y, et al. Multi-Scale residual hierarchical dense networks for single image super-resolution [J]. IEEE Access, 2019, 7: 60572-60583.

[25] WANG X T, YU K, WU S X, et al. ESRGAN: Enhanced super-resolution generative adversarial networks[C]. ECCV 2018 Workshops, 2019: 63-79.

[26] SHI W, CABALLERO J, HUSZAR F, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016: 1874-1883.

[27] MURN L, BLASI S, SMEATON A F, et al. Interpreting CNN for low complexity learned sub-pixel motion compensation in video coding[C]. 2020 IEEE International Conference on Image Processing(ICIP), 2020: 798-802.

作者简介

郑乐佳,硕士研究生,主要研究方向为视频编解码。

E-mail:zhenglejiaa@163.com

郝禄国,博士,讲师,主要研究方向为视频编解码。

E-mail:haolg@gdut.edu.cn

项颖(通信作者),博士,教授,博士生导师,主要研究方向为高光谱检测技术、AR/VR 视觉技术。

E-mail:xiangy@gdut.edu.cn

曾文彬,博士研究生,主要研究方向为深度学习,计算机视觉。

E-mail:462046548@qq.com