高技术通讯 2024 年 第 34 卷 第 7 期:726-733

doi:10.3772/j.issn.1002-0470.2024.07.006

# 基于三维卷积时空融合网络的压缩视频质量增强算法①

黄威威② 贾克斌③

(北京工业大学信息学部 北京 100124)

(北京工业大学计算智能与智能系统北京市重点实验室 北京 100124)

(先进信息网络北京实验室 北京 100124)

**摘 要** 视频数据在存储与网络传输时,通常使用标准压缩算法对原始视频进行压缩。针对压缩后视频存在压缩伪影导致视频质量下降的问题,本文提出一种基于深度学习的后处理方法提高压缩视频质量。首先,提出一种新的三维卷积时空融合网络(3D-CSTF),通过三维卷积的滤波特性提取连续视频帧之间的时空信息,并利用视频帧之间信息的强相关性来提高视频质量。其中,设计了一种用于映射和提取视频帧特征的质量增强网络(Qe-Net)。其次,将 7 个连续的视频帧送到网络进行端到端训练,利用前 3 帧和后 3 帧的信息增强当前帧。最后,在 MFQE 数据集上进行训练和测试。实验结果表明,该方法在视频质量评估标准峰值信噪比 (PSNR)上取得了良好的性能。当量化参数(QP)等于 37、32、27 和 22 时,相比压缩后的视频,PSNR 分别增加 0.82 dB、0.83 dB、0.79 dB 和 0.74 dB。

**关键词** 3D 卷积; 视频质量增强; 多帧信息; 深度学习

伴随着过去 10 年来互联网上视频数据量的迅速增长,为了在传输带宽有限条件下有效传输视频,必须对视频进行压缩以降低比特率。压缩算法(例如,H.264/AVC、H.265/HEVC)被广泛用于压缩视频数据。在视频压缩过程中,视频因为数据量化引入了各种压缩伪影,导致视频体验质量降低。因此,最大限度地减少压缩视频中的伪影,恢复压缩视频中丢失的细节,从而最终达到提高压缩视频质量的目的,已成为多媒体领域中一个重要的课题。

在过去几十年里,许多传统方法[1-4]被用于对压缩图像的质量增强研究。这些传统方法通常针对特定的压缩标准提出改进的算法,难以扩展到其他压缩方案,并且鲁棒性较差。随着神经网络的快速发展,许多基于深度学习的用于提高压缩图像[5-10]和视频[11-13]质量的方法取得了较好的结果。这些方法通过学习原始图像/视频帧与压缩图像/视频帧之间的非线性映射关系,在大量样本的训练下直接回归高质量图像和视频帧。然而,因为没有利用视频帧之间的时间信息,这些方法的改进有限。随后,Yang 等人[14]和 Guan 等人[15]提出使用多帧信息来增强压缩视频。这 2 种方法[14-15]都采用了密集的光流进行运动补偿。此外,Deng 等人[16]和 Liu 等人[17]通过可变形卷积聚合时间信息,实现了更好的增强效果。

本文设计了一个三维卷积时空融合网络(3-dimensional-convolutional spatio-temporal fusion, 3D-CSTF)来提高压缩视频的质量。其主要贡献如下:提出一种新的压缩视频质量增强方法,利用三维卷积网络的三维滤波特性,实现时空视频信息的同步提取;为了便于部署,设计的网络结构尽可能简化,主干仅由 10 个相同的卷积层结构组成,不涉及复杂的光流计算和可变形卷积的偏移计算。本文对该方

① 北京市自然科学基金(4212001)资助项目。

② 男,1997 年生,硕士生;研究方向:视频增强技术;E-mail: huangww@emails.bjut.edu.cn。

③ 通信作者,E-mail: kebinj@bjut.edu.cn。

(收稿日期:2023-07-13)

法进行了大量的实验评估,实验结果表明,该方法在提高视频质量方面具有很强的鲁棒性和有效性。

## 1 相关工作

目前,越来越多基于深度学习的压缩图像、视频质量增强的工作已经出现。根据针对的领域以及输入的视频帧数的不同,这些方法大致可以分为 3 类:即基于图像的方法、基于单帧的方法和基于多帧的方法。

### 1.1 基于图像的方法

过去十几年中,大量基于神经网络的方法被提出用于提高压缩图像的视觉质量。例如,Dong 等人[5]设计了一个紧凑高效的网络——伪影减少卷积神经网络(artifacts reduction convolutional neural network,ARCNN),它只有 4 层卷积,没有池化层和全连接层,用于衰减不同的压缩伪影。Zhang 等人[6]设计了一个去噪网络,利用残差学习和批量归一化来加快训练过程并提高去噪性能,同时拓展该网络实现单幅图像超分辨和 JPEG 图像去块功能。Yoo 等人[8]在频域中处理图像,以恢复图像压缩过程中丢失的频率分量。Chen 等人[9]结合像素域信息和小波域信息开发了一种高效的 JPEG 图像软解码方法。文献[10]中,作者提出了剩余非局部注意力学习模型残差非局部注意力网络(residual non-local attention networks,RNAN),通过保留更多的低层特征来训练深度网络,更适合于图像恢复。

### 1.2 基于单帧的方法

Dai 等人[11]提出了一种基于卷积神经网络(convolutional neural network,CNN)的高效视频编码后处理算法可变滤波器残差学习卷积神经网络(variable-filter-size residue learning convolutional neural network,VRCNN),与高效视频编码(high efficiency video coding,HEVC)基线相比,使用 VRCNN 进行后处理操作可以平均降低 4.6% 的比特率。文献[12]利用视频编码中帧内编码得到的 I 帧和预测编码得到的 P、B 帧中的信息进行训练,使训练得到的模型适合不同编码模式得到的视频帧。李子晗等人[13]提出可以同时去噪和细节补偿的细节恢复卷积神经网络(detail recovery convolutional neural network,DRCNN),该网络在补偿视频细节特征上取得了良好的效果。上述基于单帧的方法,即网络每次只输入一帧,没有考虑到帧间的相关性,限制了它们的性能。

### 1.3 基于多帧的方法

多帧质量增强(multi-frame quality enhancement,MFQE)[14]提出使用多帧信息来提高压缩视频的质量。它设计了一种基于支持向量机(support vector machine,SVM)的检测器来定位压缩视频中的峰值质量帧(peak quality frame,PQF),然后与相邻的 PQF 一起用来减少非 PQF 的压缩伪影。MFQE2.0[15]通过使用双向长短期记忆网络(bi-directional long short-term memory ,BiLSTM)存储器定位压缩视频中的 PQF,进一步改进了检测器。上述 2 种方法[14-15]为了聚合目标帧和参考帧,结合了用于运动补偿的密集光流方法。Deng 等人[16]和 Liu 等人[17]分别提出了基于可变形卷积的时空可变形融合(spatio-temporal deformable fusion,STDF)网络和多帧残差密集网络(multi-frame residual dense network,MRDN)。两个网络都通过预测目标帧和相邻参考帧的偏移场使卷积的时空采样位置变形,以捕获相关的上下文并排除噪声内容,从而提高目标帧的质量。

### 1.4 本文方法

现有的方法大部分需要额外计算连续视频帧之间的光流,由于压缩视频可能会因为各种压缩伪影而严重失真,因此估计的光流往往是不准确且不可靠的,从而导致质量增强无效。虽然基于可变形卷积网络的视频增强方案取得了较好的效果,但是由于可变形卷积的卷积核偏移的距离不同,复杂的卷积核算子导致其在工程上部署难度大幅增加。因此,本文舍弃了利用光流和可变形卷积对齐视频帧的方案,设计了一个简单的三维卷积网络,充分利用三维卷积在 3 个维度上的融合特性,同时对视频帧的空间域及时间域进行建模,并取得较好的质量增强效果。相比于其他方法,本文网络设计简单,易于部署,模型具有较强的鲁棒性,在不同量化参数(quantizer parameter,QP)压缩下的视频都有着优异的增强效果。

## 2　模型介绍

### 2.1　模型概述

视频因为压缩而产生伪影，本文的目标是通过三维卷积神经网络模型对压缩视频进行质量增强，尽可能减少视频压缩伪影。具体来说，网络分别对每个在 $t$ 时刻的压缩帧 $I_t$ 进行质量增强。

$$I_t \in R^{C\times H\times W} \tag{1}$$

其中，$R$ 表示视频帧集合，$C$ 是 $I_t$ 的通道数，$H$ 和 $W$ 分别是输入视频的高度和宽度。在 3D-CSTF 网络中，从压缩视频序列 $I=\{I_1, I_2, \cdots, I_t, I_{t+1}, \cdots, I_n\}$ 中选择 $2R$ 帧作为 $t$ 时刻增强帧 $I_t$ 的参考帧，第 1 个 $R$ 帧 $\{I_{t-R}, \cdots, I_{t-2}, I_{t-1}\}$ 是时刻 $t$ 前 $R$ 个视频帧，第 2 个 $R$ 帧 $\{I_{t+1}, I_{t+2}, \cdots, I_{t+R}\}$ 是 $t$ 时刻后 $R$ 个视频帧，增强视频帧 $\hat{I}_t \in R^{H\times W}$ 可以表示为

$$\hat{I}_t = F_\theta(I_{t-R}, I_{t-R+1}, \cdots, I_t, \cdots, I_{t+R-1}, I_{t+R}) \tag{2}$$

其中，$F_\theta$ 表示三维卷积时空融合网络（3D-CSTF）所表示的函数运算，$\theta$ 表示模型中可以学习的参数。

### 2.2　3D-CDTF 网络

本文采用三维卷积来建模输入帧之间的时间动态，三维卷积运算可以表示为

$$C(i,j,k) = \sum_{t=-\frac{T}{2}}^{\frac{T}{2}}\sum_{h=-\frac{H}{2}}^{\frac{H}{2}}\sum_{w=-\frac{W}{2}}^{\frac{W}{2}} K(t,h,w)\cdot I(i+t, j+h, k+w) \tag{3}$$

其中，$I(i,j,k)$ 表示输入特征图中的一个像素值；$C(i,j,k)$ 表示输出特征图的一个像素值，即在空间 $(i,j,k)$ 处的融合特征；$K(t,h,w)$ 表示三维卷积核中的一个权重值；$T$、$H$、$W$ 表示卷积核的尺寸。卷积核在输入特征图上滑动，对应位置的像素值与卷积核的权重值进行乘积，并求和得到输出特征图中的像素值。每个 3D 卷积层是大小为 $C_i\times C_o\times T\times H\times W$ 的滤波器，其中，$C_i$、$C_o$ 分别是输入和输出通道，$T$ 是时间维度的大小，$H$ 和 $W$ 分别是特征图的长和宽。通过 3D 卷积在 $T$、$H$、$W$ 3 个维度上的滤波特性，同时提取视频帧的时空信息。如图 1 所示，3D-CSTF 网络主要有 2 部分构成。第 1 部分中，首先将 $2R+1$ 帧视频送入第 1 层卷积进行特征预提取，第 1 层卷积的卷积核大小为 $(2R+1)\times 3\times 3$；接着，预提取特征会经过 10 个连续质量增强网络（quality enhanced network，Qe-Net）模块对时空特征进一步提取、融合与增强。每个 Qe-Net 块包含 4 层卷积和 1 个动态门控单元网络（dynamic gating unit network，DGU-Net）。动态门控单元可以自适应不同的输入，

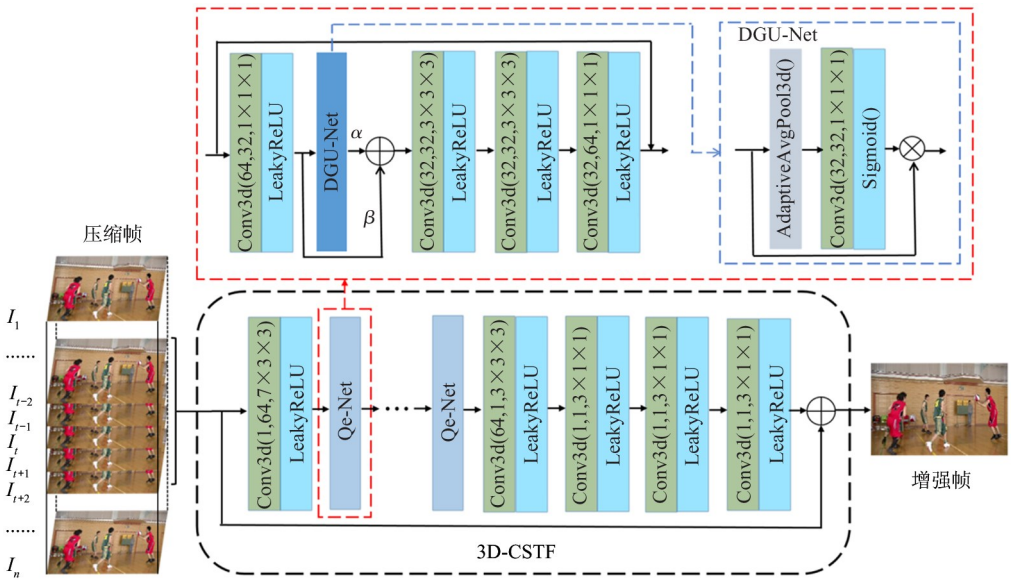


**图 1　三维卷积时空融合网络**

根据输入的特征进行动态加权,从而提高模型的适应性和表现。然而,实验发现,直接将中间特征图输入动态门控单元后输出对网络性能没有提升,本文在动态门控单元的基础上添加了 2 个可学习的参数作为门控单元的加权系数和输入的加权系数,通过跳跃连接的方式让二者相加,这样网络不仅可以学习到门控单元过滤后的信息,并且不会丢失原始信息。网络的第 2 部分使用一个 3×3×3 的卷积将输出通道压缩到 1,再使用 3 个串联的 3×1×1 卷积将经过 Qe-Net 块增强后的帧数压缩到 1 帧,在时间维度上对 Qe-Net 块提取到的特征进行融合,最后将输出与输入的 $I_t$ 帧进行全局连接得到最后的增强输出结果 $\hat{I}_t$。

### 2.3 损失函数

本文设计的网络所有层都是卷积层,每个操作(包括卷积、激活函数等)都是可导的。直接采用端到端的方式训练,总的损失函数如式(4)所示。

$$L = \sum_{i=1}^{n} (\hat{I}_t - I_t^{\text{raw}})^2 \tag{4}$$

其中,$\hat{I}_t$ 表示输入帧 $I_t$ 经过网络得到的增强帧,$I_t^{\text{raw}}$ 表示经过压缩后的输入帧 $I_t$ 对应的没有压缩的原始高质量帧,$L$ 表示 $\hat{I}_t$ 和 $I_t^{\text{raw}}$ 之间平方误差之和,即经过增强后的视频帧与原始高质量帧相应像素值间的差异。

## 3 实验结果与分析

### 3.1 数据集

实验采用 MFQE[14] 中的数据集进行训练和测试。用作真实值的 108 个视频选自数据库 Xiph(Xiph. org)[18] 和 VQEG[19],这些数据库中的视频序列分辨率范围较广:即 352×240、352×288、720×486、704×576、416×240、640×360、832×480、1 280×720、1 920×1 080 和 2 560×1 600。用于测试的 18 个视频选自视频编码联合协作小组,这些测试视频被广泛用于视频质量评估。上述所有视频都经过 H.265/HEVC 参考软件 HM16.5 压缩,压缩是在 4 个不同量化参数(QP)下进行的,QP 值分别为 22、27、32 和 37,通过在不同压缩级别下得到的视频质量增强效果来评估模型的性能。

### 3.2 实施细节

本文所采用的方法基于 PyTorch 框架实现。在训练的时候,分别从原始视频和压缩视频中随机裁剪 64×64 大小的图片作为训练样本,使用旋转或翻转来进行数据增强。采用 Adam Optimizer 优化算法进行训练,学习率设置为 0.000 5,并在整个训练过程中保持不变。模型在 4 个 QP 值下训练并测试。对于视频来说,YUV/YCbCr 空间的 Y 通道(亮度分量)包含了视频的主要信息,所以本文方法只在亮度分量上进行质量增强。视频的峰值信噪比(peak signal-to-noise ratio,PSNR)与结构相似性(structural similarity,SSIM)常用来评估视频质量的优劣,本文使用相对于 HEVC 压缩视频帧的峰值信噪比增量 ΔPSNR 和结构相似性增量 ΔSSIM 来评估质量增强的性能。

### 3.3 与现有方法比较

为了评估所提出的 3D-CSTF 网络对压缩视频增强的效果,将本文方法与各种先进的图像和视频增强方法进行了比较,包括 ARCNN[5]、DNCNN[6]、RNAN[10]、MFQE2.0[15]、STDF[16] 和 MDRN[17]。实验结果如表 1 所示。3D-CSTF 网络质量评估结果是利用输入的 7 帧视频帧得到的,即 $R = 3$,选择多个帧可以提供更大的时域范围。3 帧和 5 帧时域范围较小,大于 7 帧时,更大的时域范围提供的信息增加有限,而且会增加计算成本。为了在保持合理计算成本的同时获得足够的时域信息,本文主要实验数据在 7 帧视频帧输入下测试得到。表 1 中粗体表示最优值,斜体英文表示测试视频的名称。Class A ~ Class E 表示测试的视频分辨率分别为 2 560×1 600、1 920×1 080、832×480、416×240 和 1 280×720。本文采用的方法在 18 个测试视频上的 ΔPSNR 和 ΔSSIM 有 14 个测试视频优于所比较的方法,虽然有 4 个测试视频的结果略低于 MDRN[17] 方法,但本文方法的参数量仅占 MDRN[17] 的 46.4%。值得注意的是,在 $QP = 37$ 下,所提出的 3D-CSTF 方法的 ΔPSNR 为 0.82 dB,其性能分别超过 MDRN[17]、STDF[16]、MFQE2.0[15]、RNAN[10]、DNCNN[6] 和 ARCNN[5] 5.1%、9.3%、46.4%、86.4%、128.0% 和 228.0%。

**表 1　4 个不同 QP 的测试视频上的 ΔPSNR/ΔSSIM( $\times 10^{-2}$ )**

| *QP* | | 检测视频 | 图像增强方法 | | | 视频增强方法 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | ARCNN[5] | DNCNN[6] | RNAN[10] | MFQE2.0[15] | STDF[16] | MDRN[17] | 3D-CSTF（本文方法） |
| 37 | Class A | *Traffic* | 0.27/0.50 | 0.35/0.64 | 0.40/0.86 | 0.59/1.02 | 0.65/1.04 | 0.72/1.16 | **0.72/1.23** |
| | | *PeopleOnStreet* | 0.37/0.76 | 0.54/0.94 | 0.74/1.30 | 0.92/1.57 | 1.18/1.82 | **1.23/1.99** | 1.25/2.06 |
| | Class B | *Kimono* | 0.20/0.59 | 0.27/0.73 | 0.33/0.98 | 0.55/1.18 | 0.77/1.47 | 0.82/1.65 | **0.85/1.66** |
| | | *ParkScene* | 0.14/0.44 | 0.17/0.52 | 0.20/0.77 | 0.46/1.23 | 0.54/1.32 | **0.60/1.54** | 0.54/1.36 |
| | | *Cactus* | 0.20/0.41 | 0.28/0.53 | 0.35/0.76 | 0.50/1.00 | 0.70/1.23 | 0.67/1.30 | **0.71/1.30** |
| | | *BQTerrace* | 0.23/0.43 | 0.33/0.53 | 0.42/0.84 | 0.40/0.67 | 0.58/0.93 | 0.55/0.97 | **0.58/1.02** |
| | | *BasketballDrive* | 0.23/0.51 | 0.33/0.63 | 0.43/0.92 | 0.47/0.83 | 0.66/1.07 | 0.71/1.25 | **0.75/1.26** |
| | Class C | *RaceHorses* | 0.23/0.49 | 0.31/0.70 | 0.39/0.99 | 0.39/0.80 | 0.48/1.09 | **0.60/1.48** | 0.48/1.09 |
| | | *BQMall* | 0.28/0.69 | 0.38/0.87 | 0.45/1.15 | 0.60/1.20 | 0.90/1.61 | 0.90/1.73 | **0.95/1.80** |
| | | *PartyScene* | 0.14/0.52 | 0.22/0.69 | 0.30/0.98 | 0.36/1.18 | 0.60/1.60 | 0.55/1.66 | **0.68/2.14** |
| | | *BasketBallDrill* | 0.23/0.48 | 0.42/0.89 | 0.50/1.07 | 0.58/1.20 | 0.70/1.26 | 0.73/1.54 | **0.77/1.68** |
| | Class D | *RaceHorses* | **0.26/0.59** | **0.34/0.80** | **0.42/1.02** | **0.59/1.43** | **0.73/1.75** | **0.83/2.09** | 0.74/1.87 |
| | | *BQSquare* | **0.21/0.30** | **0.30/0.46** | **0.32/0.63** | **0.34/0.65** | **0.91/1.13** | 0.75/1.07 | **0.91/1.27** |
| | | *BlowingBubbles* | **0.16/0.46** | **0.25/0.76** | **0.31/1.08** | **0.53/1.70** | **0.68/1.96** | 0.66/2.08 | **0.71/2.20** |
| | | *BasketballPass* | **0.26/0.63** | **0.38/0.83** | **0.46/1.08** | **0.73/1.55** | **0.95/1.82** | **0.98/2.03** | 0.97/1.96 |
| | Class E | *FourPeople* | 0.40/0.56 | 0.54/0.73 | 0.70/0.97 | 0.73/0.95 | 0.92/1.07 | 0.94/1.18 | **0.97/1.38** |
| | | *Johnny* | 0.24/0.21 | 0.47/0.54 | 0.56/0.88 | 0.60/0.68 | 0.69/0.73 | 0.75/0.78 | **0.83/0.82** |
| | | *KristenAndSara* | 0.41/0.47 | 0.59/0.62 | 0.63/0.80 | 0.75/0.85 | 0.94/0.89 | 0.95/1.01 | **1.09/1.01** |
| | | 平均值 | 0.25/0.50 | 0.36/0.69 | 0.44/0.95 | 0.56/1.09 | 0.75/1.32 | 0.78/1.47 | **0.82/1.52** |
| 32 | | 平均值 | 0.19/0.17 | 0.33/0.41 | 0.41/0.62 | 0.52/0.68 | 0.73/0.87 | 0.81/1.02 | **0.83/1.06** |
| 27 | | 平均值 | 0.16/0.09 | 0.33/0.26 | 0.39/0.30 | 0.49/0.42 | 0.67/0.53 | **0.83/0.72** | 0.79/0.68 |
| 22 | | 平均值 | 0.13/0.04 | 0.27/0.14 | 0.28/0.16 | 0.46/0.27 | 0.57/0.30 | 0.70/0.40 | **0.74/0.44** |

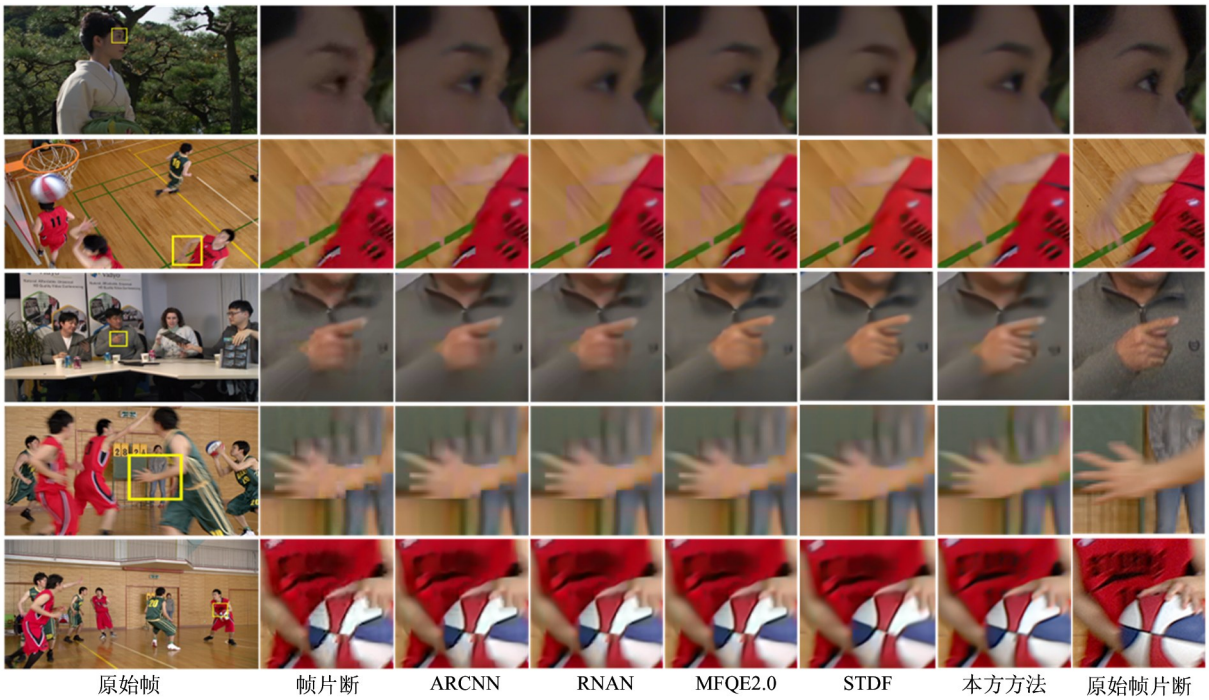


**图 2　QP37 时的质量增强效果**

通过全面比较4种不同QPs的ΔPSNR值,本文提出的模型表现出较好的鲁棒性。具体地说,不同QP下得到的ΔPSNR最大差异仅为0.09 dB,MDRN[17]、STDF[16]、MFQE2.0[15]、RNAN[10]、DNCNN[6]和ARCNN[5]方法最大差异分别为0.13 dB、0.18 dB、0.1 dB、0.16 dB、0.09 dB和0.13 dB。

### 3.4 质量增强效果

为了对所提出的方法进行定性评估,图2中给出了部分可视化示例。结果表明,传统的图像质量增强方法虽然可以减少压缩伪影,但往往会导致图像细节丢失。相反,所提出的多帧视频质量增强方法利用参考帧得到了更好的增强结果。例如,对于图2第3行的*FourPeople*视频,经过压缩后,手指变得模糊,手指之间的缝隙几乎看不见。虽然本研究中对比的方法减少了部分伪影,但仍然无法恢复手指的清晰度。相比之下,本文提出的3D-CSTF网络有效地消除了大量伪影,并基本上恢复了手指之间的缝隙,从而获得更清晰的图像。

### 3.5 减少质量波动

质量波动是评估增强视频整体质量的关键指标,因为大的质量波动会破坏视频帧的时间一致性并降低主观视频质量。为了评估本文提出的方法对质量波动的影响,图3绘制了2个视频序列(*BQSquare*视频的第50~100帧和*PartyScene*视频的第100~150帧)的PSNR折线图。如图所示,在HEVC标准压缩之后,原始视频的PSNR表现出显著的波动。为了提高压缩效率,压缩后的视频帧一般分为关键帧I帧(独立编码,不依赖于其他帧进行解码,PSNR值较高)、前向预测帧P帧(依赖于当前帧之前的帧编解码,PSNR值较低)和双向预测帧B帧(利用当前帧前后帧编解码,PSNR最低)。所以可以发现,经过HEVC标准压缩后,由于I、P、B帧的压缩效率不同,PSNR波动较大。使用本文提出的方法增强压缩视频可以有效地缓解视频帧之间的质量波动。通过数据分析,可以观察到通过本文的方法获得的相邻帧之间最大PSNR差异显著低于原始压缩视频和单帧增强方法。这些结果突出了本文方法在提升视频时间一致性和增强主观质量方面的潜力。

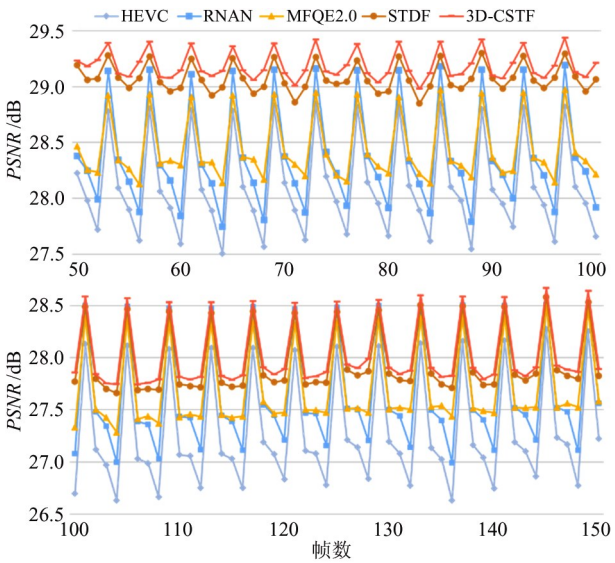


**图3 测试视频的PSNR折线图**

## 4 消融实验

### 4.1 Qe-Net模块

为了证明3D-CSTF网络中主体结构Qe-Net模块的良好性能,以不加动态门控单元的Qe-Net模块作为比较的base模块,并分别以下面的1个结构替换以验证base模块的效果。结构1:将base模块替换成只有1层3×3×3卷积的结构。结构2:将base模块替换成3个卷积核分别为1×1×1、3×3×3和1×1×1卷积层串联的结构。结构3:将base模块替换成5个卷积核分别为1×1×1、3×3×3、3×3×3、3×3×3和1×1×1串联的结构。为了公平比较,实验保持不同base块的参数量和计算量基本不变,结果如表2所示。可以发现,在相同的训练策略下,Qe-Net模块在相对较低的参数量与计算量的情况下比上述3种方案得到的Δ*PSNR*分别高了23.8%、9.86%和4.00%。

**表2 不同3D卷积结果**

| Qe-Net | ΔPSNR/ΔSSIM | 参数量/M | FLOPs/M |
|---|---|---|---|
| 结构1 | 0.63/1.23 | 669 | 307.05 |
| 结构2 | 0.71/1.31 | 622 | 284.86 |
| 结构3 | 0.75/1.34 | 616 | 282.15 |
| base | 0.78/1.41 | 601 | 275.11 |

### 4.2 动态门控单元

动态门控单元不同的信息分配方式结果如表3

所示。当动态门控单元直接串联在 Qe-Net 模块中，即当 $\alpha = 1$ 和 $\beta = 0$ 时，模型的性能显著降低。然而，通过在动态门控单元中添加残差并包含可学习参数 $\beta$ 时，模型的性能略有提高。当 $\alpha$ 和 $\beta$ 同时可以学习时，模型的性能显著提升，与 base 模型相比，当 *QP* 等于 32 时，*PSNR* 增加 0.08 dB。值得注意的是，动态门控单元的参数量仅为 $1.05 \times 10^4$，FLOPs 仅为 0.147 M。

**表 3 动态门控信息分配**

| *QP* | 分配方案 | | | |
|---|---|---|---|---|
| | base | base + ($\alpha = 1$) + ($\beta = 0$) | base + ($\alpha = 1$) + $\beta$ | base + $\alpha + \beta$ |
| 37 | 0.78/1.41 | 0.73/1.23 | 0.78/1.42 | 0.82/1.52 |
| 32 | 0.75/0.94 | 0.71/0.89 | 0.76/0.95 | 0.83/1.06 |
| 27 | 0.75/0.65 | 0.66/0.58 | 0.75/0.66 | 0.79/0.68 |
| 22 | 0.72/0.43 | 0.63/0.37 | 0.73/0.43 | 0.74/0.44 |

### 4.3 时间信息融合

相邻的视频帧之间通常存在一定的相关性，通过融合多帧视频，可以捕捉到运动信息、场景变化以及相邻帧之间的关系。为了证明本文方法在融合时间信息方面（即连续的多帧视频间的信息）的有效性，通过改变输入网络的视频帧的数量来进行实验，同时保持网络结构基本不变，然后重新训练模型。实验结果如表 4 所示，其中 *R*1、*R*2 和 *R*3 分别表示输入网络的视频帧数量为 3 帧、5 帧和 7 帧。以 *QP* = 37 时为例，使用 7 帧视频作为输入比使用 5 帧视频作为输入时 *PSNR* 增加了 6.4%，比使用 3 帧视频作为输入时 *PSNR* 增加了 20.6%。

**表 4 不同数量的输入帧结果**

| *QP* | 输入方案 | | |
|---|---|---|---|
| | 3D-CSTF(*R*1) | 3D-CSTF(*R*2) | 3D-CSTF(*R*3) |
| 37 | 0.68/1.19 | 0.77/1.40 | 0.82/1.52 |
| 32 | 0.70/0.89 | 0.79/1.00 | 0.83/1.06 |
| 27 | 0.66/0.55 | 0.70/0.60 | 0.79/0.68 |
| 22 | 0.62/0.34 | 0.68/0.39 | 0.74/0.44 |

## 5 结 论

本文设计了一种新的用于压缩视频质量增强的多帧网络——三维卷积时空融合网络 3D-CSTF，区别于常用的使用光流进行运动估计和使用可变形卷积对齐空间特征的多帧网络，本文使用 3D 卷积结构实现在连续视频帧上同步提取融合时空信息。大量实验表明，本文提出的 3D-CSTF 方法明显提高了压缩视频的质量，在基准数据库上实现了先进的性能；与其他方法相比，其鲁棒性更强，Δ*PSNR* 和 Δ*SSIM* 更高，质量波动更小。

本文目前的工作仍局限于采用 3D 卷积，尽管具有较小的参数量，但计算复杂度仍然较高，推理速度会受到影响，从而影响了实时应用需求。未来的工作将集中于优化网络架构以提高推理速度。

**参考文献**


[1] FOI A, KATKOVNIK V, EGIAZARIAN K. Pointwise shape-adaptive DCT for high-quality denoising and deblocking of grayscale and color images[J]. IEEE Transactions on Image Processing, 2007,16(5):1395-1411.

[2] SIKORAT. Low complexity shape-adaptive DCT for coding of arbitrarily shaped image segments[J]. Signal Processing Image Communication, 1995,7(4-6):381-395.

[3] JANCSARY J, NOWOZIN S, ROTHERC. Loss-specific training of non-parametric image restoration models: a new state of the art[C]//The 12th European Conference on Computer Vision. Florence, Italy: Springer, 2012: 112-125.

[4] CHIOU Y W, YEH C H, KANG L W, et al. Efficient image/video deblocking via sparse representation[C]// 2012 Visual Communications and Image Processing. San Diego, USA: IEEE, 2012:1-6.

[5] DONG C, DENG Y, LOYC C, et al. Compression artifacts reduction by a deep convolutional network[C]// Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015:576-584.

[6] ZHANG K, ZUO W, CHEN Y, et al. Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising[J]. IEEE transactions on image processing, 2017,26(7):3142-3155.

[7] LI K, BARE B, YAN B. An efficient deep convolutional neural networks model for compressed image deblocking [C]//2017 IEEE International Conference on Multimedia and Expo. Hong Kong, China:IEEE, 2017:1320-1325.

[8] YOO J, LEE S, KWAK N. Image restoration by estimating frequency distribution of local patches[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018:

6684-6692.
[9] CHEN H, HE X, QING L, et al. DPW-SDNet:dual pixel-wavelet domain deep CNNs for soft decoding of JPEG-compressed images[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City, USA: IEEE, 2018:711-720.
[10] ZHANG Y, LI K, LI K, et al. Residual non-local attention networks for image restoration[C]// The 7th International Conference on Learning Representations. New Orleans, USA: ICLR, 2019:1-18.
[11] DAI Y, LIU D, WU F. A convolutional neural network approach for post-processing in HEVC intra coding[C]// The 23rd International Conference on Multi Media Modeling. Reykjavik, Iceland: Springer, 2017:28-39.
[12] YANG R, XU M, WANG Z. Decoder-side HEVC quality enhancement with scalable convolutional neural network[C]//2017 IEEE International Conference on Multimedia and Expo. Hong Kong, China: IEEE, 2017:817-822.
[13] 李子晗, 邵笑, 张佩云. 基于细节还原卷积神经网络的压缩视频质量增强技术研究[J]. 南京信息工程大学学报(自然科学版),2023,15(3):274-285.
[14] YANG R, XU M, WANG Z, et al. Multi-frame quality enhancement for compressed video[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018:6664-6673.
[15] GUAN Z, XING Q, XU M, et al. MFQE 2.0: a new approach for multi-frame quality enhancement on compressed video[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019,43(3):949-963.
[16] DENG J, WANG L, PU S, et al. Spatio-temporal deformable convolution for compressed video quality enhancement[C]// Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI Press, 2020:10696-10703.
[17] LIU J, ZHOU M, XIAO M. Deformable convolution dense network for compressed video quality enhancement[C]// The 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore: IEEE, 2022:1930-1934.
[18] Xiph. org. Xiph. orgvideo test media[EB/OL]. [2023-07-05]. https://media. xiph. org/video/derf/.
[19] VQEG. VQEG video datasets and organizations[EB/OL]. [2023-07-05]. https://www. its. bldrdoc. gov/vqeg/video-datasets-and-organizations. aspx.

# Compressed video quality enhancement algorithm based on 3D convolutional spatio-temporal fusion network

HUANG Weiwei, JIA Kebin
(Faculty of Information Technology, Beijing University of Technology, Beijing 100124)
(Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing University of Technology, Beijing 100124)
(Beijing Laboratory of Advanced Information Networks, Beijing 100124)

**Abstract**

Standard compression algorithms are typically used to compress video data for storage and transmission over networks. However, compressed video can have compression artifacts that degrade quality. To address this problem, a post-processing method based on deep learning is proposed. Firstly, a novel 3-dimensional convolutional spatio-temporal fusion(3D-CSTF) network is designed, which extracts the temporal information between consecutive video frames through the filtering characteristics of the 3D convolution kernel in three dimensions, and utilizes the strong correlation of the information between video frames to enhance the video quality. Among it, a quality enhanced network (Qe-Net) is designed for mapping and extracting video frame features. Secondly, seven consecutive video frames are sent to the network for end-to-end training and the current frame is enhanced by using the information of the previous and last three frames. Finally, training and testing are carried out on the MFQEv2 dataset. Experimental results demonstrate that this method achieves excellent performance in terms of the video quality assessment standard PSNR. When the quantization parameter (QP) are equal to 37, 32, 27 and 22, the PSNR can be increased by 0.82 dB, 0.83 dB, 0.79 dB and 0.74 dB, respectively.

**Key words**: 3-dimensional convolution, video quality enhancement, multi-frame information, deep learning