

联合视口预测和超分辨率重建的实时 VR 全景视频流式传输方法

陈晓雷, 曹宝宁, 卢禹冰, 张鹏程

(兰州理工大学电气工程与信息工程学院 兰州 730050)
(chenxl703@lut.edu.cn)

摘要: 现有基于视口预测的 VR 全景视频流式传输方法尚未有效地考虑 VR 视频的时间维度信息, 并且忽略了客户端在提高视频重建质量上应发挥的作用. 为了进一步提高传输系统的综合性能及用户体验质量, 提出一种联合视口预测和超分辨率重建的实时 VR 全景视频流式传输方法. 在服务器端提出时间非局部注意力模块, 并将其嵌入 GhostNet 中对 VR 全景视频进行全局上下文建模, 捕获视频内容特征在时间维度上的长距离依赖; 在客户端提出轻量级 VR 全景视频超分辨率重建模型, 对来自服务器端预测视口内的次要内容进行质量增强及投影失真优化. 在 VR 全景视频用户头部运动数据集上的实验结果表明, 所提方法的平均视口预测精度和平均带宽占用分别为 95.6% 和 52.9%, 与 5 种代表性的传输方法相比, 该方法能够获得更高的视口预测精度和更低的带宽占用, 同时具有良好的视频重建质量和较低的计算资源消耗.

关键词: 360°视频; 实时虚拟现实流; 视口预测; 超分辨率重建

中图分类号: TP391.41

DOI: 10.3724/SP.J.1089.2024.19994

Live VR Panoramic Video Streaming Method Combining Viewport Prediction and Super-Resolution

Chen Xiaolei, Cao Baoning, Lu Yubing, and Zhang Pengcheng

(School of Electrical Engineering and Information Engineering, Lanzhou University of Technology, Lanzhou 730050)

Abstract: The existing VR panoramic video streaming methods based on viewport prediction have not effectively considered the temporal dimension information of VR video, and ignored the role that the client end should play in improving the video reconstruction quality. To further improve the comprehensive performance of the streaming system and the user quality of experience, a live VR panoramic video streaming method that combines viewport prediction and super-resolution reconstruction is proposed. At the server end, the method captures the long-distance dependency in the temporal dimension of video content features by embedding the proposed temporal non-local attention module TNAM into GhostNet to model the global context of VR panoramic video; At the client end, the proposed lightweight VR panoramic video super-resolution reconstruction model LVRSR is used to enhance the quality and optimize the projection distortion of the secondary content within the predicted viewport from the server end. The experimental results on the VR panoramic video user head motion dataset show that the average viewport prediction accuracy and average bandwidth usage of the method are 95.6% and 52.9%, respectively. Compared with five representative streaming methods, the method can achieve higher viewport prediction accuracy and lower bandwidth usage, while having good video reconstruction quality and low computational resource consumption.

收稿日期: 2022-10-08; 修回日期: 2023-07-01. 基金项目: 国家自然科学基金(61967012); 甘肃省科技计划(20JR5RA459).
陈晓雷(1979—), 男, 博士, 副教授, 硕士生导师, 主要研究方向为人工智能、计算机视觉、虚拟现实; 曹宝宁(1997—), 男, 硕士研究生, 主要研究方向为虚拟现实、图像处理、人工智能; 卢禹冰(1997—), 男, 硕士, 主要研究方向为图像处理、姿态估计; 张鹏程(1996—), 男, 硕士研究生, 主要研究方向为虚拟现实、图像显著性检测.

Key words: 360° video; live VR streaming; viewport prediction; super-resolution reconstruction

随着5G移动通信网络 and 多媒体技术的快速发展,VR全景视频(又称360°视频^[1-2])在消费市场越来越受欢迎,成为当前视频处理和传输中的研究热点^[3-6]。使用者佩戴头戴式显示器(head mounted display, HMD)观看VR全景视频时,在同一位置能够通过改变观察方向浏览四周全部场景,获得传统二维视频不能提供的沉浸感、真实感和交互感。然而,VR全景视频具有高分辨率(high resolution, HR)特性,为了保证用户体验,现有条件下,4K分辨率是最低要求,未来可能要求分辨率达到6K,8K甚至12K;同时,由于VR全景视频覆盖的视口高达360°,因此与视口小于50°的传统二维视频相比,在相同分辨率和视频长度下,VR全景视频的数据量比二维视频大5倍以上。面对如此巨大的数据量,在现有无线网络带宽条件下,要传输一个拥有足够分辨率并且覆盖完整视口的VR全景视频会占用过多的带宽资源^[7],导致视频缓冲延迟,影响VR全景视频点播的用户体验;对于实时VR全景视频传输(即VR全景视频直播),这一问题将更加严重。

用户佩戴HMD观看VR全景视频时,呈现在他们眼前的视频场景将随着头部位置的改变而改变,即在某一时刻,用户只能看到对应视口部分的视频内容,因此以相同码率传输所有的360°场景是不必要的。为了节省网络带宽并保证VR全景视频的传输质量,近年来,基于视口预测的VR全景视频流式传输方法^[8-23]成为研究热点。这些方法的基本思路是:在服务器端,基于用户观看轨迹或VR全景视频内容预测用户未来的视口,并将VR全景视频帧分为若干视频块(Tiles);然后将视口内的Tiles以高码率向客户端传输,视口外的Tiles以低码率进行传输甚至不传输。

现有方法虽然在一定程度上实现了VR全景视频的有效传输,但是忽略了客户端在提高视频重建质量上应发挥的作用。从原理上考虑,将服务器端的视口预测和客户端的超分辨率重建^[24-26]有机联合,能进一步提高整个传输系统的综合性能。因此,本文提出一种联合视口预测和超分辨率重建的实时VR全景视频流式传输方法。首先提出一种时间非局部注意力模块(temporal non-local attention module, TNAM),将其嵌入GhostNet^[27]中,构建提取VR全景视频特征的网络模型(GhostNet with TNAM, T-GhostNet),可以对VR全景视频进行全局上下文建模,捕获视频内容特征在时间维度

上的长距离依赖,提升视口预测精度;同时使用GRU-ECA模型^[23]感知用户视口轨迹,将两者结合作为视口预测模型以预测用户的未来视口。然后提出一种用于客户端的轻量级VR全景视频超分辨率重建模型(lightweight network for VR panoramic video super-resolution, LVRSR),利用重参数化^[28]确保在高速重建的同时保持良好的视频重建质量,并使用Shuffle结构^[29]控制模型大小;针对VR全景视频的等距柱状投影(equirectangular projection, ERP)失真问题,提出多纬度带重建方法。最后基于所提出的视口预测和超分辨率重建模型,提出一种用于VR全景视频直播的实时VR全景视频流式传输方法,具有高视口预测精度和低带宽占用特性,能获得良好的视频重建质量且计算资源消耗低。

针对本文方法进行了充分的实验,并与当前代表性方法进行对比,结果表明,该方法的综合性能优于对比方法,能更好地支持实时VR全景视频流式传输。

1 相关工作

1.1 VR全景视频视口预测

视口预测是决定VR全景视频传输系统性能的核心环节,预测视口与用户实际视口的吻合度越高,用户体验越好;反之则越差。目前,视口预测方法可以分为2大类:基于用户观看轨迹的方法和基于VR全景视频内容的方法。

(1) 基于用户观看轨迹的方法利用用户的头部运动特征预测未来视口。van Damme等^[8]使用经典的监督学习方法,利用不同用户的历史头部方向预测未来视口,并根据用户类型和视频类型调整算法的复杂性以增强预测性能;但这种跨用户方法需要其他用户的历史观看信息,只适合于点播场景。Qian等^[9]分别使用基于窗口的平均值、逻辑回归和加权逻辑回归等简单方法,利用头部运动数据预测未来视口。Heyse等^[10]提出一种基于上下文的方法,利用头部旋转信息估计用户的未来视口。Xu等^[11]提出一种概率预测模型,利用用户头部方向的概率分布预测未来视口。Hou等^[12]提出一种多层长短期记忆(long short-term memory, LSTM)模型学习一般的头部运动模式,并根据过去的轨迹预测未来的视口。Yang等^[13]将分层贝叶斯推理集成到LSTM中进行头部轨迹预测,从过

去、当前和估计的未来信息中捕获时间相关性. Jiang 等^[14]将用户头部原始旋转量转换为正弦值来减小偏航方向的预测误差; 由于用户观看行为的内在多变性, 上述方法仅利用用户观看轨迹很难获得较高的视口预测精度.

(2) 基于 VR 全景视频内容的方法仅利用视频内容特征, 或同时利用用户头部运动特征和视频内容特征进行视口预测. Xu 等^[15]提出一种基于深度强化学习的方法, 以更好地模拟用户对视频内容的注意力. Chen 等^[16]使用卷积神经网络(convolutional neural networks, CNN)提取 VR 全景视频内容的特征, 使用 LSTM 作为位置相关数据的编码来预测用户的未来视口. 为了解决视频投影失真引起的加权失效问题, Li 等^[17]提出使用球形卷积提取 VR 全景视频的显著时空特征, 并利用有限的视口反馈信息预测用户未来的视口: 这些方法由于需要采集大量用户长期观看 VR 全景视频的数据, 通常仅适用于点播场景. 针对 VR 全景视频直播面临的独特实时挑战, Feng 等^[18-21]先后通过更改深度神经网络模型的训练/推理过程、结合光流和高斯模型以及使用强化学习跟踪检测对象等方法, 消除对用户长期观看数据的依赖, 使用实时观看数据实现对用户视口的在线预测. 在 Feng 等^[20]工作的基础上, Peng 等^[22]结合视口预测和基于 Tiles 的自适应比特率算法优化 VR 全景视频的传输: 以上方法由于会消耗过多的服务器资源, 在实际应用中的系统部署负担较重. 因此, Chen 等^[23]使用剪枝的轻量级网络代替传统 CNN 以减少服务器负担, 并使用注意力增强模块提升预测精度. 此外, 基于这些方法构建的 VR 全景视频传输方法的带宽占用仍有进一步优化的空间.

1.2 VR 全景视频超分辨率重建

目前, 针对 VR 全景视频的超分辨率重建方法研究较少. Liu 等^[24]提出单帧和多帧联合网络, 可以由低分辨率(low resolution, LR)VR 输入视频恢复出 HR 输出视频, 利用像素级帧间一致性, 采用变形卷积消除目标帧特征映射与相邻帧之间的运动差异; 但是, 该方法不能与服务端端的视口预测方法相结合构成 VR 全景视频传输系统, 因为在基于视口预测的传输系统中, 需要超分辨率重建的 Tiles 并不总是处于连续帧中的相同位置. Deng 等^[25]提出基于尺度自适应网络的超分辨率重建方法, 采用不同的尺度因子恢复具有不同内容复杂度的像素, 显著地提高了 VR 全景视频的超分辨率重建性能; 但该方法的复杂度较高, 无法满足 VR 全景视频直播的实时性要求. Dasari 等^[26]将较短的视频片段放在服务器端训练微型模型, 然后将微型模型与视频片段一起传输到客户端并进行超分辨率重建, 在保持超分辨率重建性能的同时可以降低计算成本; 但该方法没有对 VR 全景视频的投影失真特点进行优化, 影响重建质量.

2 本文方法

2.1 总体框架

本文方法分为服务器端和客户端 2 部分, 其总体框架如图 1 所示. 其中, 服务器端负责视口预测, 客户端负责超分辨率重建.

在服务器端, 首先实时收集直播中的视频片段, 等待用户观看相应的视频片段后反馈得到视口轨迹, 并基于用户反馈对训练视频帧进行标记; 考虑实时性要求, 对每个片段的视频帧和相应的

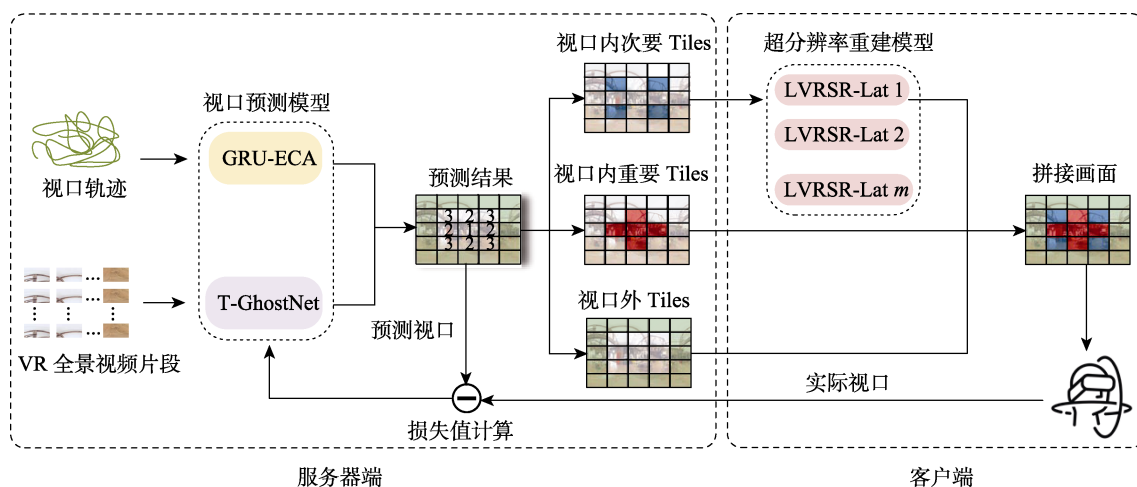


图 1 本文方法总体框架

视口轨迹进行二次下采样处理, 然后将每帧均匀地分割成小的子图像; 在视口预测模型中, 利用 T-GhostNet 模型分析用户对视频内容的喜好, 并利用 GRU-ECA 模型^[23]感知用户视口轨迹, 将这 2 个模型的结果融合产生最初的预测视口; 最后将预测视口分为重要 Tiles 和次要 Tiles, 将重要 Tiles 以高码率直接发送给客户端, 次要 Tiles 进行下采样后得到 LR Tiles, 并与预测视口外 Tiles 一起以低码率发送给客户端。

在客户端, 首先使用 LVRSR 模型对来自服务器端的次要 LR Tiles 进行多纬度带重建, 得到 HR Tiles; 然后将 HR Tiles、来自服务器端的重要 Tiles 和视口外 Tiles 拼接, 构成客户端完整的预测画面; 将其解码和渲染后呈现在 HMD 等用户观看设备上; 当用户观看完当前视频片段后, 将预测视口与实际视口在服务器端进行比较并计算损失值, 利用损失值相应地更新 T-GhostNet 模型和 GRU-ECA 模型的权重, 以实时匹配用户当前的观看行为。

按照服务器端和客户端的工作原理, 本文方法通过收集视频和用户视口轨迹, 依次对每个 VR 全景视频片段进行视口预测和超分辨率重建, 直到所有视频片段传输完成。

2.2 视口预测模型

为了在实时 VR 全景视频流式传输中获得高精度的视口预测, 本文使用 T-GhostNet 和 GRU-ECA 的混合模型进行预测, 对 2 个模型预测结果对应的

视口内 Tiles 取并集作为预测视口, 即同时保留重合位置和非重合位置的 Tiles. 其中, GRU-ECA 模型^[23]通过增强视口轨迹序列之间的隐含关系, 在几乎不增加延迟开销的同时可以提高用户视口的预测精度. 下面仅详细描述 T-GhostNet 模型的工作原理。

2.2.1 TNAM

已有的视口预测工作常采用单一的 CNN 提取 VR 全景视频特征, 忽略了 VR 全景视频中的时间维度信息. 针对这一问题, Feng 等^[21]和兰诚栋等^[1]采用循环神经网络或三维卷积(three-dimensional convolution, Conv3D)神经网络提取视频的时间特征; 但使用这类局部操作捕获长距离依赖需要累积很多层的网络, 导致学习效率过低, 而非局部操作通过直接计算 2 个位置之间的关系就能快速地捕获长距离依赖. 因此, 本文将非局部注意力 NLNet^[30]或 GCNet^[31]嵌入 CNN 对 VR 全景视频进行全局上下文建模, 捕获视频内容特征在时间维度上的长距离依赖. 然而, NLNet 巨大的参数量和计算量会消耗过多的计算资源, GCNet 虽然进一步实现了轻量化, 但只能捕获空间维度上的长距离依赖. 因此, 本文提出一种即插即用的 TNAM, 与其他非局部注意力模块相比, 该模块可以捕获 VR 全景视频内容特征在时间维度上的长距离依赖, 并保持轻量化的特点。

TNAM 结构如图 2 所示, 其工作原理分为 3 个步骤。

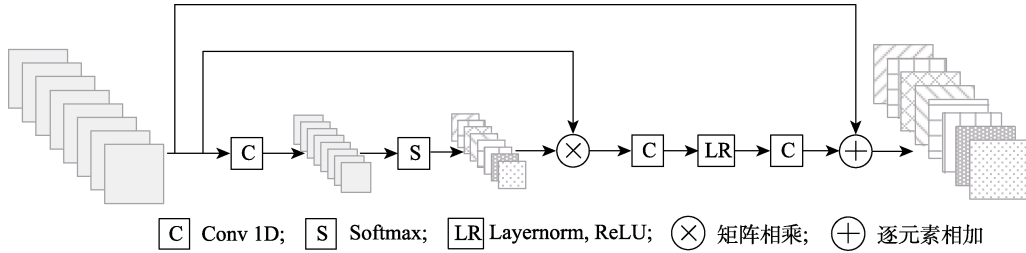


图 2 TNAM 结构

Step1. 全局上下文建模. 设输入的数据为 $\mathbf{X} \in \mathbb{R}^{(b \times c \times h \times w)}$; 其中, $b = n \times t$, n 和 t 分别表示每个 VR 全景视频帧包含的 Tiles 总数和每段视频的总帧数, h, w, c 分别表示每帧图像长、宽和通道数. 将 $\mathbf{X}$ 进行维度变换, 得到 $\mathbf{X}_1 = \mathbf{X}_2 \in \mathbb{R}^{(n \times (c \times h \times w) \times t)}$. $\mathbf{X}_1$ 首先经过一个一维卷积后, 将所有通道和空间特征映射到一个点中, 减少参数量; 然后经过指数映射(Softmax)放大差异输出为 $\mathbf{Y}_1 = \text{Softmax}(\mathbf{X}_1 * \mathbf{f}_1)$. 其中, $\mathbf{Y}_1 \in \mathbb{R}^{(n \times 1 \times t)}$ 和 $\mathbf{f}_1 \in \mathbb{R}^{((c \times h \times w) \times 1 \times 1)}$ 为一维卷积的卷积核大小; 最后, 将 $\mathbf{Y}_1$ 的第 2 维和第 3 维进行维度交换后与 $\mathbf{X}_2$ 相乘, 初步得到含

有时间维度全局上下文信息的特征图 $\mathbf{Y}_2 = \mathbf{X}_2^T \mathbf{Y}_1$. 其中, $\mathbf{Y}_2 \in \mathbb{R}^{(n \times (c \times h \times w) \times t)}$.

Step2. 捕获 VR 全景视频内容特征的通道间依赖. 为了进一步获取视频内容特征通道间的重要程度并减少计算资源消耗, 先对 $\mathbf{Y}_2$ 进行一维卷积以压缩通道数; 然后使用层归一化(layer normalization, LN)降低优化难度, 且作为正则提高泛化性; 最后经过激活函数 ReLU, 使用一维卷积复原通道数得到输出

$$\mathbf{Y}_3 = \text{ReLU}(\text{LN}(\mathbf{Y}_2 * \mathbf{f}_2) * \mathbf{f}_3).$$

其中, $\mathbf{Y}_3 \in \mathbb{R}^{(n \times (c \times h \times w) \times 1)}$; $\mathbf{f}_2 \in \mathbb{R}^{(c \times 0.25 \times 1)}$ 和 $\mathbf{f}_3 \in \mathbb{R}^{(0.25 \times c \times 1)}$

分别表示第 1 个卷积和第 2 个卷积的卷积核大小。

Step3. 特征融合. 将 Y_3 经过广播机制 R 后与 X_1 逐元素相加, 聚合全局上下文特征到每个像素位置; 然后进行维度变换, 得到最终输出 $Y = X_1 + R(Y_3)$. 其中, $Y \in \mathbb{R}^{(b \times c \times h \times w)}$.

2.2.2 T-GhostNet

GhostNet 是一种新型的轻量级 CNN, 以更加简易的卷积方式从更少的参数中生成更多的特征图, 在保持较高的识别性能的同时降低卷积层的计算成本. 本文将 TNAM 嵌入 GhostNet 中作为 T-GhostNet. 除了提取空间特征之外, T-GhostNet 还可以对 VR 全景视频进行全局上下文建模, 捕获视频内容特征在时间维度上的长距离依赖. 根据

本文视口预测方法的实际需求, 只输出“感兴趣”和“不感兴趣”2 个类别, 为了避免不必要的计算资源浪费, 还对 GhostNet 的通道进行缩减, 如图 3a 所示, 它由二维卷积(two-dimensional convolution, Conv2D), G-Bneck (Ghost BottleNeck)、嵌入 SENet^[32]的 G-Bneck 和全连接(full connection, FC)层构成. 根据已有的设计经验, 将注意力模块嵌入不同的位置会影响最终的预测性能, 因此, 本文将 TNAM 分别嵌入通道缩减的 GhostNet 卷积头、G-Bneck(包含嵌入 SENet 的 G-Bneck)和卷积尾后进行了对比实验, 然后选择最优结果作为最终的嵌入位置, 图 3b 所示为在卷积头后嵌入 TNAM 的 T-GhostNet 结构图.

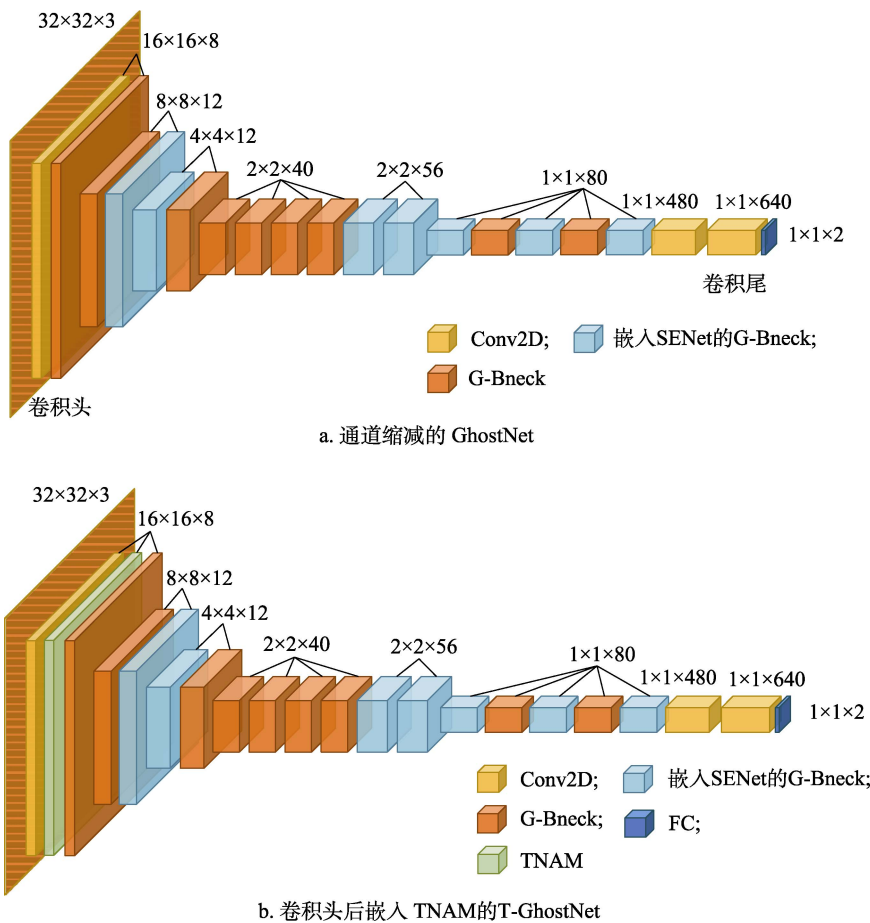


图 3 通道缩减的 GhostNet 与 T-GhostNet 结构图

2.3 预测视口 Tiles 分类

受限于客户端设备的算力和 VR 全景视频的 HR 特性, 现有方法^[24-26]难以保证在流畅直播的同时实现对预测视口内所有 Tiles 的实时超分辨率重建. 为此, 本文将预测视口内的 Tiles 分为重要 Tiles 和次要 Tiles 这 2 大类, 其中, 重要 Tiles 以高码率直接发送给客户端; 次要 Tiles 进行下采样后,

以低码率传输到客户端进行放大因子为 2 的超分辨率重建. Tiles 分类应考虑如下因素.

(1) 依据 Chakareski 等^[33]和 Chen 等^[34]的实验结果, 将预测视口内的 Tiles 分为多种关注度等级: 中心 Tiles 用户关注的可能性最高, 赋予最高等级; 水平和垂直相邻 Tiles 用户关注的可能性次高, 赋予中等等级; 对角相邻 Tiles 的关注可能性最低,

赋予最低等级。

(2) 考虑超分辨率重建的实时性及整个传输方法的时延限制(即单个视频片段时长 T , 本文中实验设置为 2 s), 根据可用超分辨率重建时间 t_{sr} 和客户端设备可用算力 s_{client} , 确定每个视频片段中可以实时 LR 传输的 Tiles 最大数量

$$N = \max f(t_{sr}, s_{client}) \quad (1)$$

其中, t_{sr} 需要满足 $t_{vp} + t_{sr} + t_{other} \leq T$, t_{vp} 表示视口预测时间, t_{other} 表示 VR 全景视频采集时间、Tiles 传输时间和 Tiles 拼接时间的总和。

(3) 将用户关注度可能性最低的 N 个 Tiles 作为次要 Tiles, 剩余的 Tiles 作为重要 Tiles。

图 1 中包含单个视频段的 Tiles 分类示例。综

合考虑上述因素, 首先根据用户关注特性, 将预测视口内的 Tiles 分为 3 种等级(图 1 中用数字代表用户关注度, 数字越小则关注度越高); 然后按照式(1)计算得到 $N=4$; 最后将图 1 中的 4 个蓝色 Tiles 作为次要 Tiles, 其余的 5 个红色 Tiles 作为重要 Tiles。

2.4 超分辨率重建模型

为了应对 VR 全景视频超分辨率重建的实时挑战, 本文提出 LVRSR 模型, 其构造过程如图 4 所示。首先在现有重参数化方法上增加 1×1 卷积分支, 形成重参数化(re-parameterization, RP)模块; 然后在 RP 模块的基础上构建重参数化 Shuffle (RP shuffle, RPS)块; 最后为 LVRSR 设计一个简洁的主干网络。

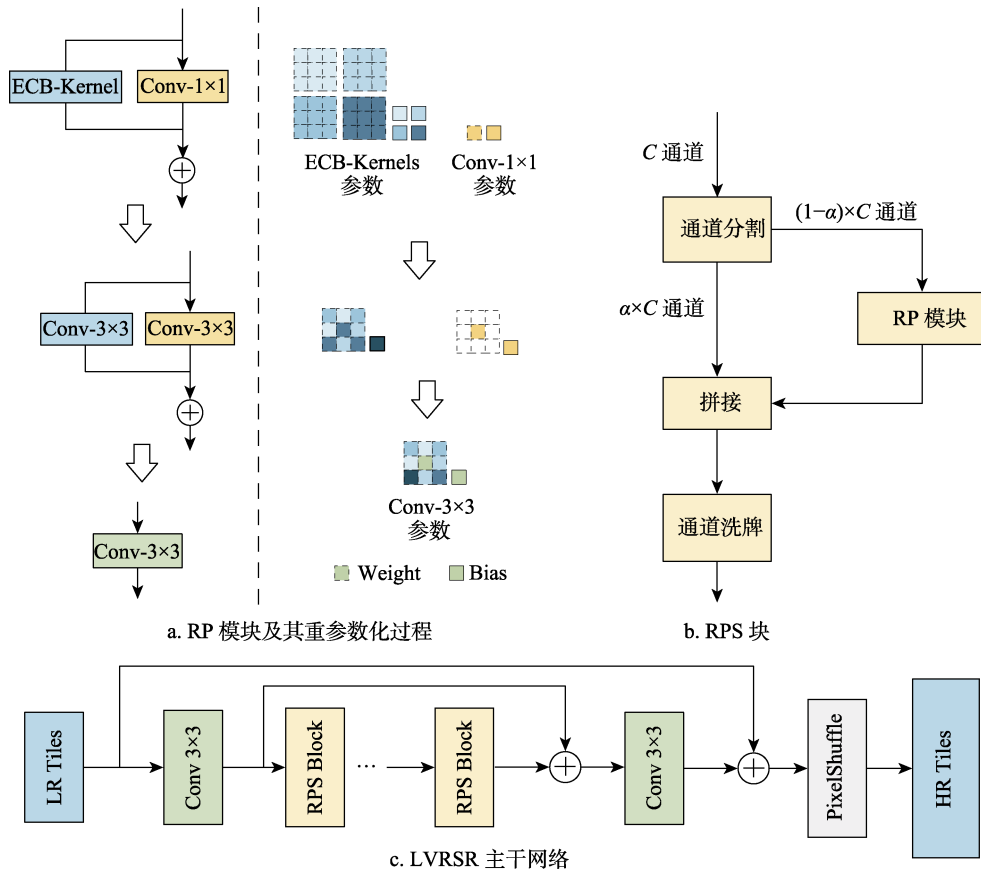


图 4 LVRSR 模型构造过程

2.4.1 RP 模块

在推理时, RP 方法既能拥有多分支网络带来的丰富特征, 又能大幅提升推理速度, 训练时采用多分支网络, 推理时将其合并为单分支网络。然而, 现有 VR 全景视频超分辨率重建方法没有考虑到 1×1 卷积的 RP, 而 1×1 卷积可以融合不同通道同一位置像素的特征, 丰富特征提取的尺度。因

此, 本文在 ECB(edge-oriented convolution block)^[35]的基础上增加 1×1 卷积分支, 构成 RP 模块。

RP 模块的结构及对应的 RP 过程如图 4a 所示。训练时, ECB 分支与 1×1 卷积分支并行, 此时的输出可表示为

$$F = F_{ECB} + F_{1 \times 1} = W_{ECB} * X + B_{ECB} + W_{1 \times 1} * X + B_{1 \times 1}.$$

其中, F_{ECB} , W_{ECB} 和 B_{ECB} 分别表示 ECB 的输出、

权重和偏置; $F_{1 \times 1}$, $W_{1 \times 1}$ 和 $B_{1 \times 1}$ 分别表示 1×1 卷积的输出、权重和偏置; F 表示训练阶段 RP 模块的输出.

1×1 卷积的 RP 过程可表示为

$$F'_{1 \times 1} = W'_{1 \times 1} * X + B'_{1 \times 1} = P(W_{1 \times 1} * X) + B_{1 \times 1}.$$

其中, $F'_{1 \times 1}$, $W'_{1 \times 1}$ 和 $B'_{1 \times 1}$ 分别表示 RP 1×1 卷积的输出、权重和偏置; P 表示在 1×1 卷积核周围填充一圈 0, 成为 3×3 卷积.

模型推理时的网络结构合并为一个 3×3 卷积, 其合并方式为

$$F' = W' * X + B' = (W'_{ECB} + W'_{1 \times 1}) * X + B'_{ECB} + B'_{1 \times 1}.$$

其中, W'_{ECB} 和 B'_{ECB} 分别表示 RP ECB 的权重和偏置; F' , W' 和 B' 分别表示推理阶段 RP 模块的输出、权重和偏置.

2.4.2 RPS 块

Shuffle 结构可以根据分割系数 α , 将一个具有 C 通道的张量沿通道维度分割成 2 个分支. 其中, 使用 $(1-\alpha) \times C$ 个通道的分支进行深度卷积或点卷积; 而另一个分支处于空闲状态, 用于传递剩余的 $\alpha \times C$ 个通道信息. 将 2 个分支连接起来, 进行通道洗牌后构造块的输出, 使其具有与输入相同的通道数. 随着 α 的增加可以堆叠更多的网络层, 同时控制网络参数量. 除此之外, 根据 Liu 等^[36]的实验, 通道分割还可以起到通道注意的作用, 通过反向传播在特定通道上提供更复杂的权值, 这种结构不仅计算效率高, 而且能够在超分辨率重建中提升重建质量. 因此, 本文将 RP 模块替换卷积分支以构造 RPS 块, 如图 4b 所示.

2.4.3 LVRSR 主干网络

由于复杂的拓扑结构(如密集连接)会严重影响推理速度^[29], 因此本文使用简洁的拓扑结构设计 LVRSR 主干网络, 如图 4c 所示. 该网络使用一个 3×3 卷积作为卷积头、一个 3×3 卷积作为卷积尾和一系列 RPS 块将 LR 视频帧映射到高维空间进行特征提取, 最后使用 PixelShuffle 层重建出 HR 视频帧. 此外, 仅使用 2 个残差结构来丰富特征表示, 不会过于影响模型的推理速度.

2.5 多纬度带重建

目前, 常采用 ERP 投影将 VR 全景视频转换成二维视频进行处理和传输, 然而这种投影方式会带来不同程度的失真, 纬度越高像素被水平拉伸程度越高, 像素密度越小, 而这对于二维平面卷积的平移不变性具有破坏作用. 通常的解决方式是在训练过程中使用具有修正失真权重的损失函

数, 如加权均方误差(weighted mean square error, WMSE)^[37]. 但是, 原始 WMSE 只能针对整幅 VR 全景视频帧进行训练, 而本文超分辨率重建的对象是 Tiles, 因此本文提出多纬度带重建方法, 通过对训练过程进行修改优化失真影响.

首先将 VR 全景视频帧分割为 m 个纬度带的 Tiles, 与视口预测中的 Tiles 分块方案保持一致; 然后使用对应纬度带的 Tiles 和损失函数进行训练, 以获得 m 个具有不同纬度带权重的 LVRSR; 最后在推理时, 对不同纬度带的 Tiles 使用对应权重的 LVRSR 进行超分辨率重建. 损失函数公式为

$$WMSE_k = \frac{\sum_{i=0}^{X-1} \sum_{j=0}^{Y-1} (y_k(i, j) - y'_k(i, j))^2 w_k(i, j)}{\sum_{i=0}^{X-1} \sum_{j=0}^{Y-1} w_k(i, j)},$$

其中, $y_k(i, j)$ 和 $y'_k(i, j)$ 分别表示分辨率为 $X \times Y$ 的测试视频帧, 以及参考视频帧上第 k 个纬度带 (i, j)

位置处的采样值; $w_k(i, j) = \cos\left(\left(j - \frac{N}{2} + \frac{1}{2}\right) \frac{\pi}{N}\right)$ 表示第 k 个纬度带中从等长方形到单位球面的面积比例因子.

3 实验与结果分析

3.1 数据集

选择 VR 全景视频用户头部运动数据集(简称为 VR 全景头部数据集)^[38]的不同数据, 用于训练和测试视口预测模型. 该数据集是一个由 48 个用户观看 5 个类别、18 个 VR 全景视频组成的头部运动数据集, 不仅提供了用户头部运动数据记录, 还包括实验中使用的视频. 为了评估本文方法的性能, 从该数据集中选取 8 个具有代表性的测试视频进行实验, 包括体育、表演、脱口秀和纪实视频, 这些视频的相关信息如表 1 所示. 其中, 视频 3 和视频 8 由多个摄像机捕获, 视频 1、视频 4 和视频 5 中的背景是静态的, 其余视频中的动态背景是由背景本身的运动、摄像机的移动或多个摄像机之间的切换引起的.

对经典数据集 DIV2K^[39]进行数据增强, 用于训练超分辨率重建模型 LVRSR, 即在 DIV2K 基础上增加 VR 全景图像数据集 ODISR^[25]. ODISR 数据集中收集了 1 000 幅高质量的 VR 全景图像, 包含城市景观、自然景观、人类活动和展览等不同类型的內容.

表1 测试视频

序号	视频名称	视频类型	视频背景	摄像机数量
1	Cooking	表演	静态	1
2	Football	体育	动态	1
3	Fighting	体育	动态	2
4	Anitta	表演	静态	1
5	RioVR	脱口秀	静态	1
6	Front	体育	动态	1
7	Falluja	纪录片	动态	1
8	Rhinos	纪录片	动态	2

本文选择 VR 全景头部数据集用于整体传输方法的测试。

3.2 实验环境

本文实验在 GPU 模式下进行,并使用 CUDA 加速深度学习处理。服务器端计算机的配置为:32 GB 内存, Intel(R)Core(TM)i3-9100F CPU, NVIDIA GTX1080Ti GPU, 64 位 Ubuntu 16.04 操作系统;客户端计算机的配置为:16 GB 内存, AMD R5-4600H CPU, NVIDIA GTX1650 GPU, 64 位 Windows 10 操作系统。实验选用的脚本语言为 Python, 并安装 Pytorch 深度学习框架。

3.3 实验设置

对于 T-GhostNet 的输入, 首先将 VR 全景视频分为时长 2 s 的视频段; 然后对其依次进行抽帧、降采样和 5×5 均匀分割, 得到 200 个 32×32 大小的 Tiles, 即 Batch_size=200; 为了满足 VR 全景视频直播的实时性要求, 将 Epoch 设置为 10。对于 GRU-ECA 的输入, 将用户视口轨迹进行与视频段相同的采样预处理; 然后采用当前视频段和相应视口轨迹段进行视口预测模型训练; 最后在下一个视频段和当前视口轨迹段上进行模型测试, 分别预测下一个视频段和相应视口轨迹段的用户视口。

对于 LVRSR, 采用 HR 图像和 2 倍 Bicubic 下采样图像进行训练, Epoch 设置为 1 000。将 DIV2K 数据集的图像随机裁剪为 224×224 像素大小, 并对其分别进行水平和垂直翻转, 以增强数据。将 ODISR 数据集的图像均分为 5 个纬度带、 80×200 像素大小, 将不同纬度带的图像使用对应的损失函数进行训练。模型训练完成后, 在 VR 全景头部数据集上进行测试。

3.4 评价指标

采用 6 种评价指标, 对本文方法的性能进行全面评估。

(1) 预测精度。通过计算视频中全部正确预测帧的数量在所有帧数量中所占的比例来获取预测

精度, 当实际视口完全被预测视口覆盖的情况下, 视为预测正确;

(2) 带宽占用。根据文献[20]的计算方法, 忽略所有以低码率传输的 Tiles(视口外的 Tiles 和视口内的次要 Tiles)以便于计算, 将高码率传输的重要 Tiles 在所有 Tiles 中所占的比例定义为带宽占用;

(3) 预测时间。从 VR 全景视频采集完成到获得单个视频片段的视口预测结果的时间差, 将其作为视口预测时间;

(4) 计算资源消耗。采用模型参数量和乘-累加运算(multiply-add operations, MACs)来衡量模型计算资源的消耗情况;

(5) 超分辨率重建客观质量评价。对于超分辨率重建的 Tiles, 使用加权球面均匀峰值信噪比(weighted to spherically uniform PSNR, WS-PSNR)^[40]和加权球面均匀结构相似性(weighted to spherically uniform SSIM, WS-SSIM)^[41]对其进行客观质量评价;

(6) 超分辨率重建速度。使用每秒重建的帧数衡量超分辨率重建的速度。

3.5 超参数确定

为了使本文方法获得最佳综合性能, 需要调整超参数。下面讨论其中一些重要的超参数对实验结果的影响。

3.5.1 TNAM 添加位置

表 2 所示为视口预测模型在不同 TNAM 嵌入位置下对预测性能的影响。可以看出, 在卷积头和 G-Benck 后嵌入均可获得较高的预测精度, 其中, 嵌入卷积头后取得了最少的预测时间, 然而在 G-Benck 后嵌入参数量和 MACs 增长较多; 在卷积尾后嵌入虽然获得了最少的参数量和 MACs, 但预测精度下降。因此, 为了在预测时间、预测精度、参数量和 MACs 之间建立良好的平衡, 最终选择将 TNAM 嵌入 GhostNet 的卷积头后。

表2 视口预测模型在不同 TNAM 嵌入位置下的预测性能对比

位置	预测时间/s	预测精度/%	参数量	MACs
卷积头后	0.468	95.6	1.353M	379.812M
G-Benck 后	0.535	95.0	1.368M	380.825M
卷积尾后	0.470	95.2	1.350M	379.601M

注: 粗体表示最优值。

3.5.2 RPS 块分割系数

为了选择 RPS 块的最佳通道分割系数, 采用 3 种不同的方案进行实验, 如表 3 所示。参考已有的设计经验, 没有选择对每层的 RPS 块都进行分

割,而是对其中部分 RPS 块进行分割,并将这些 RPS 块的数量记为 S ,把未进行分割的 RPS 块的数量记为 K ,通道数记为 C .为了对比不同分割系数下 LVRSR 的性能表现,统一设置 $K=2$, $S=5$, $C=8$ (称为 LVRSR-K2S5C8) 方案进行实验.可以看出,当分割系数 $\alpha=0.7$ 时, LVRSR 获得了最差的重建性能,这是因为采用大分割系数后只有少数通道进行非线性映射,导致难以提取到 VR 全景视频帧更多的特征;减少分割系数时均获得了较高的 WS-PSNR 和 WS-SSIM,但与 $\alpha=0.4$ 相比, $\alpha=0.2$ 时的重建性能提升不大,反而增加了模型参数量与 MACs,会消耗更多的客户端计算资源.因此,最终选择 $\alpha=0.4$ 作为 RPS 块的分割系数.

表 3 LVRSR-K2S5C8 在不同 RPS 块分割系数下的超分辨率重建性能对比

α	WS-PSNR/dB	WS-SSIM	参数量	MACs
0.7	36.22	0.955 0	0.001 764M	88.510M
0.4	36.34	0.955 8	0.002 739M	137.432M
0.2	36.36	0.955 8	0.003 244M	162.771M

注. 粗体表示最优值.

表 4 视口预测模型的消融实验结果对比

方法	预测时间/ms	预测精度/%	带宽占用/%	参数量	MACs
Baseline	461	94.4	[55.3, 64.9]	1.344M	379.249M
Baseline+TNAM	468	95.6	[55.8, 64.2]	1.353M	379.812M
NLNet	470	95.6	[55.2, 65.1]	1.380M	382.112M
LSTM	505	91.1	[54.9, 64.5]	1.352M	455.348M
Conv3D	473	91.4	[56.3, 65.4]	1.347M	467.722M
Baseline+LVRSR-K2S5C8		95.6	[48.2, 56.6]		

注. Baseline+LVRSR-K2S5C8 为本文整个传输方法; 粗体表示最优值.

将 ECBSR^[34]作为基准模型(Baseline-SR), 并采用与 LVRSR 相同的主干网络, 超分辨率重建 LVRSR-K2S5C8 模型的消融实验结果如表 5 所示.可以看出, 加入 RP 1×1 卷积(RP-Conv 1×1)后, WS-PSNR 和 WS-SSIM 均有提升, 说明融合同一位置不同像素的特征有助于提升重建质量; 使用 Shuffle 结构后, 由于通道分割的作用, WS-PSNR 和 WS-SSIM 显著增加, 而参数量和 MACs 减少; 最后使用多纬度带重建后, 由于其对 VR 全景视频投影失真的优化作用, 因此 WS-PSNR 和 WS-SSIM 进一步提升.

以上实验结果表明, 本文所提的各个模块和方法都是有效的和必要的. 进一步, 对传输过程中

3.6 消融实验

为了证明本文方法每个模块和训练方法的有效性, 分别对视口预测模型和超分辨率重建模型 LVRSR-K2S5C8 进行消融实验. 此外, 还将 TNAM 分别替换为 NLNet, LSTM 和 Conv3D, 对比基准模型在分别加入这 4 种时间特征提取模块后的性能表现.

将 LightLive^[23]作为基准模型(Baseline), 视口预测模型的消融实验结果如表 4 所示. 可以看出, 加入 TNAM 后, 参数量和 MACs 略有增长, 预测时间几乎没有变化, 预测精度提升 1.3 个百分点, 但此时带宽占用并没有明显变化, 而对预测视口内的次要 Tiles 进行下采样传输并使用 LVRSR 进行超分辨率重建后, 带宽占用下降了 13.6 个百分点; 将 TNAM 替换为其他模块后, 使用同为非局部注意力的 NLNet 和 TNAM 在预测精度上取得了相同的分数, 但 NLNet 由于自身算法的复杂性, 导致其参数量和 MACs 大幅增加, 而使用 LSTM 和 Conv3D 后, 视口预测模型难以优化使得对预测精度产生了负面影响. 因此, 本文提出的 TNAM 整体性能优于其他时间特征提取模块.

表 5 LVRSR-K2S5C8 的消融实验结果对比

方法	WS-PSNR/dB	WS-SSIM	参数量	MACs
Baseline-SR	36.23	0.955 4	0.002 748M	137.884M
Baseline-SR+RP-Conv 1×1	36.29	0.955 7	0.002 748M	137.884M
Baseline-SR+RPS 块	36.34	0.956 0	0.002 739M	137.432M
Baseline-SR+多纬度带重建	36.39	0.956 2	0.002 739M	137.432M

注. 粗体表示最优值.

的延迟情况进行测试.

图 5 所示为本文方法对 8 个测试视频每个分段的平均预测时间累积分布(cumulative distribution

function, CDF)结果.可以看出,大多数视频段的处理时间在 300~700 ms,所有视频段的预测时间都小于 800 ms,远远小于视频分段时长 2 s.实验结果证明,本文方法可以在当前视频片段的播放过程中实时完成视口预测,并且有更充裕的时间在客户端完成超分辨率重建,而不会延迟到下一个片段,能够支持流畅的实时VR全景视频直播.

3.7 与当前代表性方法的对比实验

为了证明本文方法的优越性,将本文方法与当前实时VR全景视频流式传输领域和超分辨率重建领域的代表性方法分别进行定性和定量对比.

在各个测试视频上,不同实时VR全景视频流式传输方法的预测精度对比如图6所示.可以看出,本文方法在绝大多数视频上都具有最高的预测精度,整体精度保持在90%以上,而对比方法的性能并不稳定,在不同视频中的预测准确率差异很大,最大差异可达49.6个百分点;与LightLive相比,本文方法的平均预测精度提高1.3个百分点,其中,该方法大幅提升了视频2、视频3和视频8

的预测精度,说明使用TNAM可以在一定程度上抵抗动态背景和多场景切换所带来的干扰;但是,用户在观看背景高度动态的VR全景视频时,头部运动具有更高的任意性,与其他视频相比,本文方法对视频6和视频7的预测精度仍处于较低水平.

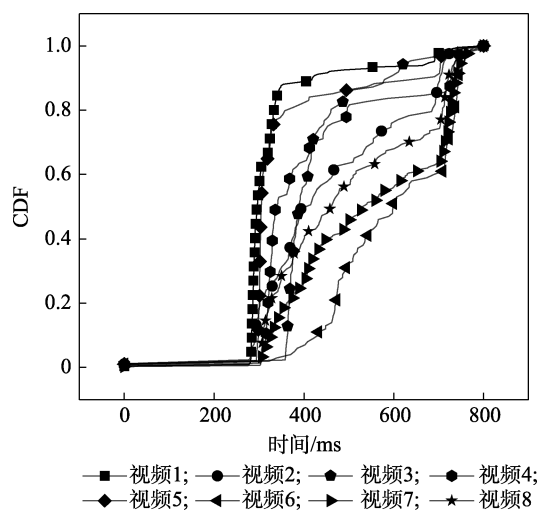


图5 8个测试视频的CDF

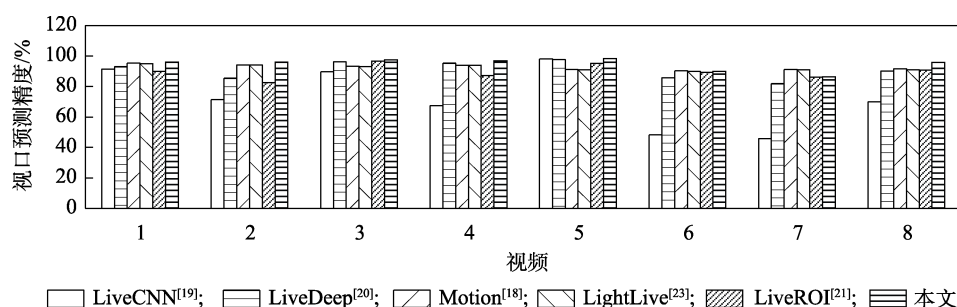


图6 6种方法在8个测试视频下的视口预测精度对比

图7所示为6种方法的带宽占用情况的箱线图.可以看出, Motion^[18]会导致不同视频和用户的带宽使用存在巨大差异; LiveROI^[21], LiveCNN^[19], LiveDeep^[20]和 LightLive^[23]在大多数视频上的带宽占用较为稳定,但仍处在较高水平;为了尽可能减少带宽占用,本文在客户端部署了重建速度最快的LVR SR-K2S5C8超分辨率重建模型,加入该模型后,本文方法的带宽占用大幅降低,与LiveROI相比,平均带宽占用下降5.9个百分点.值得注意的是,随着客户端设备的算力增加,带宽占用可进一步减少,说明本文方法在减少带宽占用方面具有巨大潜力.

在2倍放大因子下,将本文提出的LVR SR模型与其他轻量级超分辨率重建方法的客观指标进行对比,结果如表6所示.为了直观地展示LVR SR的优势,将其按照模型参数量扩充至3个模型规格(LVR SR-K2S5C8, LVR SR-K3S6C12和LVR SR-

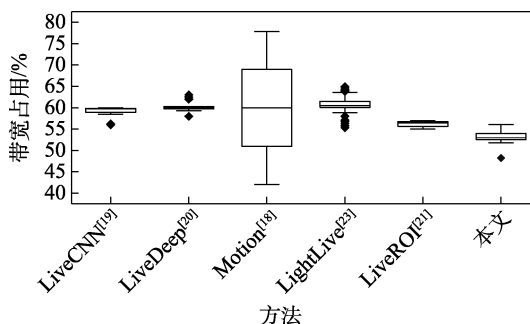


图7 6种方法的带宽占用对比

K5S10C16);然后在同一模型规格区间内,与大于LVR SR模型参数数量的超分辨率重建方法^[42-46]进行对比.可以看出,在不同的模型规格区间内,采用本文方法, LVR SR与ECBSR的重建速度均保持领先水平,特别是LVR SR-K2S5C8的重建速度远高于的其他方法,可以大幅减少VR全景视频传输的

带宽占用; 在绝大多数情况下, LVRSR 取得了最高的 WS-PSNR 和 WS-SSIM 分数, 同时具有最少的参数量和 MACs. 其中, 由于模型大小的限制, LVRSR-K5S10C16 在 WS-PSNR 方面略微低于 360SR^[46], 但参数量和 MACs 明显少于 360SR, 重建速度超出

360SR 方法 2 倍; 在重建速度接近的情况下, 与对比方法相比, LVRSR-K3S6C12 和 LVRSR-K5S10C16 的 WS-PSNR 和 WS-SSIM 优势较为微弱, 说明仅依靠简单地增加 S , K 和 C , 并不能明显地提升模型的客观重建质量.

表 6 2 倍放大因子下 LVRSR 与不同超分辨率重建方法的客观评价指标对比

规格	方法	重建速度/(帧·s ⁻¹)	WS-PSNR/dB	WS-SSIM	参数量	MACs
小	Bicubic		34.32	0.926 8		
	SRCNN ^[42]	47.6	35.53	0.949 3	0.024 281M	2 874M
	FSRCNN ^[43]	60.5	36.28	0.955 7	0.012 809M	1 326M
	ESPCN ^[44]	55.5	35.65	0.952 1	0.021 284M	1 068M
	LVRSR-K2S5C8	78.9	36.39	0.956 2	0.002 739M	137M
中	RDN-S ^[45]	33.3	37.06	0.956 5	0.011 697M	598M
	ECBSR-M4C16 ^[35]	71.4	37.24	0.956 9	0.010 200M	340M
	LVRSR-K3S6C12	71.6	37.28	0.957 7	0.007 258M	364M
大	ECBSR-M10C16 ^[35]	66.7	37.46	0.959 0	0.024 300M	5 600M
	360SR ^[46]	21.3	37.52	0.958 2	0.586 379M	31 799M
	LVRSR-K5S10C16	66.9	37.51	0.959 1	0.021 636M	1 086M

注: 粗体表示最优值.

图 8 所示为 LVRSR 与不同超分辨率重建方法对视频 1 和视频 3 的主观重建质量对比. 可以看出, 与同一模型规格区间的其他方法相比, 由于本文所提的各个模块和方法的作用, LVRSR 获得了更好的视觉质量, 如纹理更清晰、边缘更锐利和多余光晕更少, 即使采用最小复杂度的模型也能达到

良好的重建质量, 并且随着模型复杂度的增加, LVRSR 的重建效果也更接近原始 HR VR 全景视频帧. 实验结果表明, 将视口预测与超分辨率重建结合是可行的, 特别是采用本文提出的 LVRSR 之后, 可以在大幅减少传输 VR 全景视频所需带宽的同时, 保持客户端较高的视频重建质量.

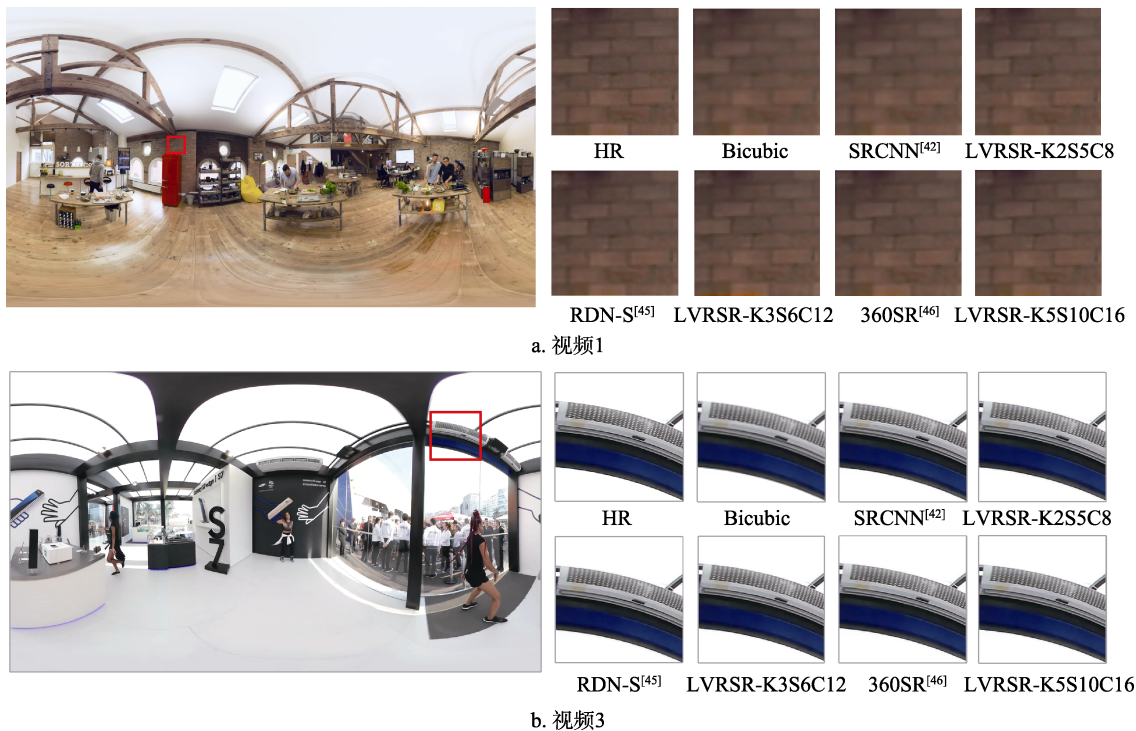


图 8 2 倍放大因子下 LVRSR 与不同超分辨率重建方法的主观重建质量对比

4 结 语

本文首先提出 TNAM 并嵌入 GhostNet 中对 VR 全景视频进行全局上下文建模,捕获视频内容特征在时间维度上的长距离依赖;然后提出用于 VR 全景视频的超分辨率重建网络 LVRSR,在实现轻量化的同时针对 ERP 投影失真进行了优化;在此基础上,提出一种基于视口预测和超分辨率重建的实时 VR 全景视频流式传输方法。大量实验结果表明,本文方法在拥有更高的视口预测精度和更低的传输带宽占用率前提下,能获得更好的视频质量和更低的计算资源消耗,整体性能优于对比方法。然而,本文方法对于拥有动态背景视频的视口预测精度提升仍然有限,未来,将结合 VR 全景视频内容的语义信息提高此类视频的视口预测精度。针对在执行速度接近的情况下,本文提出的 LVRSR 的精度优势不明显问题,今后可通过使用球形卷积代替普通卷积优化 VR 全景视频投影失真来解决。

参考文献(References):

- [1] Lan Chengdong, Rao Yingjie, Song Caixia, *et al.* Adaptive streaming of stereoscopic panoramic video based on reinforcement learning[J]. *Journal of Electronics & Information Technology*, 2022, 44(4): 1461-1468(in Chinese)
(兰诚栋, 饶迎节, 宋彩霞, 等. 基于强化学习的立体全景视频自适应流[J]. *电子与信息学报*, 2022, 44(4): 1461-1468)
- [2] Ye Chengying, Li Jianwei, Chen Sixi. Research progress of VR panoramic video transmission[J]. *Application Research of Computers*, 2022, 39(6): 1601-1607+1621(in Chinese)
(叶成英, 李建微, 陈思喜. VR 全景视频传输研究进展[J]. *计算机应用研究*, 2022, 39(6): 1601-1607+1621)
- [3] Chiariotti F. A survey on 360-degree video: coding, quality of experience and streaming[J]. *Computer Communications*, 2021, 177: 133-155
- [4] Yaqoob A, Bi T, Muntean G M. A survey on adaptive 360° video streaming: solutions, challenges and opportunities[J]. *IEEE Communications Surveys & Tutorials*, 2020, 22(4): 2801-2838
- [5] Han Shengqian, Lou Han, Wang Junlai. Traffic evaluation and modeling for spatio-temporal slicing based VR panoramic video transmission[J]. *Journal on Communications*, 2022, 43(7): 102-112(in Chinese)
(韩圣千, 娄函, 王君来. VR 全景视频时空切片传输的流量评估与建模[J]. *通信学报*, 2022, 43(7): 102-112)
- [6] Ruan J J, Xie D L. Networked VR: state of the art, solutions, and challenges[J]. *Electronics*, 2021, 10(2): Article No.166
- [7] Anwar M S, Wang J, Ullah A, *et al.* Measuring quality of experience for 360-degree videos in virtual reality[J]. *Science China Information Sciences*, 2020, 63(10): Article No.202301
- [8] van Damme S, Vega M T, de Turck F. Machine learning based content-agnostic viewport prediction for 360-degree video[J]. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2022, 18(2): Article No.50
- [9] Qian F, Ji L S, Han B, *et al.* Optimizing 360 video delivery over cellular networks[C] // *Proceedings of the 5th Workshop on All Things Cellular: Operations, Applications and Challenges*. New York: ACM Press, 2016: 1-6
- [10] Heyse J, Vega M T, de Backere F, *et al.* Contextual bandit learning-based viewport prediction for 360 video[C] // *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces*. Los Alamitos: IEEE Computer Society Press, 2019: 972-973
- [11] Xu Z M, Zhang X G, Zhang K, *et al.* Probabilistic viewport adaptive streaming for 360-degree videos[C] // *Proceedings of the IEEE International Symposium on Circuits and Systems*. Los Alamitos: IEEE Computer Society Press, 2018: 1-5
- [12] Hou X S, Dey S, Zhang J Z, *et al.* Predictive view generation to enable mobile 360-degree and VR experiences[C] // *Proceedings of the Morning Workshop on Virtual Reality and Augmented Reality Network*. New York: ACM Press, 2018: 20-26
- [13] Yang L, Xu M, Guo Y C, *et al.* Hierarchical Bayesian LSTM for head trajectory prediction on omnidirectional images[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 44(11): 7563-7580
- [14] Jiang X L, Naas S A, Chiang Y H, *et al.* SVP: sinusoidal viewport prediction for 360-degree video streaming[J]. *IEEE Access*, 2020, 8: 164471-164481
- [15] Xu M, Song Y H, Wang J Y, *et al.* Predicting head movement in panoramic video: a deep reinforcement learning approach[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, 41(11): 2693-2708
- [16] Chen X W, Kasgari A T Z, Saad W. Deep learning for content-based personalized viewport prediction of 360-degree VR videos[J]. *IEEE Networking Letters*, 2020, 2(2): 81-84
- [17] Li J, Han L, Zhang C, *et al.* Spherical convolution empowered FoV prediction in 360-degree video multicast with limited FoV feedback[OL]. [2022-10-08]. <https://arxiv.org/abs/2201.12525>
- [18] Feng X L, Swaminathan V, Wei S. Viewport prediction for live 360-degree mobile video streaming using user-content hybrid motion tracking[J]. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2019, 3(2): Article No.43
- [19] Feng X L, Bao Z Y, Wei S. Exploring CNN-based viewport prediction for live virtual reality streaming[C] // *Proceedings of the IEEE International Conference on Artificial Intelligence and Virtual Reality*. Los Alamitos: IEEE Computer Society Press, 2019: 183-186
- [20] Feng X L, Liu Y, Wei S. LiveDeep: online viewport prediction for live virtual reality streaming using lifelong deep learning[C] // *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces*. Los Alamitos: IEEE Computer Society Press, 2020: 800-808
- [21] Feng X L, Li W T, Wei S. LiveROI: region of interest analysis for viewport prediction in live mobile virtual reality streaming[C] // *Proceedings of the 12th ACM Multimedia Systems Conference*. New York: ACM Press, 2021: 132-145
- [22] Peng S, Hu J L, Xiao H, *et al.* Viewport-driven adaptive 360°

- live streaming optimization framework[J]. *Journal of Networking and Network Applications*, 2022, 1(4): 139-149
- [23] Chen X L, Cao B N, Ahmad I. Lightweight neural network-based viewport prediction for live VR streaming in wireless video sensor network[J]. *Mobile Information Systems*, 2021, 2021: Article No.8501990
- [24] Liu H Y, Ruan Z B, Fang C W, *et al.* A single frame and multi-frame joint network for 360-degree panorama video super-resolution[OL]. [2022-10-08]. <https://arxiv.org/pdf/2008.10320>
- [25] Deng X, Wang H, Xu M, *et al.* LAU-Net: latitude adaptive up-scaling network for omnidirectional image super-resolution[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 9185-9194
- [26] Dasari M, Bhattacharya A, Vargas S, *et al.* Streaming 360-degree videos using super-resolution[C] //Proceedings of the IEEE Conference on Computer Communications. Los Alamitos: IEEE Computer Society Press, 2020: 1977-1986
- [27] Han K, Wang Y H, Tian Q, *et al.* GhostNet: more features from cheap operations[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 1577-1586
- [28] Ding X H, Zhang X Y, Ma N N, *et al.* RepVGG: making VGG-style ConvNets great again[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 13728-13737
- [29] Ma N N, Zhang X Y, Zheng H T, *et al.* ShuffleNet V2: practical guidelines for efficient CNN architecture design[C] //Proceedings of the 15th European Conference on Computer Vision. Heidelberg: Springer, 2018: 122-138
- [30] Wang X L, Girshick R, Gupta A, *et al.* Non-local neural networks[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 7794-7803
- [31] Cao Y, Xu J R, Lin S, *et al.* GCNet: non-local networks meet squeeze-excitation networks and beyond[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. Los Alamitos: IEEE Computer Society Press, 2019: 1971-1980
- [32] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 7132-7141
- [33] Chakareski J, Aksu R, Corbillon X, *et al.* Viewport-driven rate-distortion optimized 360° video streaming[C] //Proceedings of the IEEE International Conference on Communications. Los Alamitos: IEEE Computer Society Press, 2018: 1-7
- [34] Chen X L, Wu D, Ahmad I. Optimized viewport-adaptive 360-degree video streaming[J]. *CAA Transactions on Intelligence Technology*, 2021, 6(3): 347-359
- [35] Zhang X D, Zeng H, Zhang L. Edge-oriented convolution block for real-time super resolution on mobile devices[C] //Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM Press, 2021: 4034-4043
- [36] Liu X, Li Y A, Fromm J, *et al.* SplitSR: an end-to-end approach to super-resolution on mobile devices[J]. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2021, 5(1): Article No.25
- [37] Tran H T T, Ngoc N P, Bui C M, *et al.* An evaluation of quality metrics for 360 videos[C] //Proceedings of the 9th International Conference on Ubiquitous and Future Networks. Los Alamitos: IEEE Computer Society Press, 2017: 7-11
- [38] Wu C L, Tan Z H, Wang Z, *et al.* A dataset for exploring user behaviors in VR spherical video streaming[C] //Proceedings of the 8th ACM on Multimedia Systems Conference. New York: ACM Press, 2017: 193-198
- [39] Agustsson E, Timofte R. NTIRE 2017 challenge on single image super-resolution: dataset and study[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Los Alamitos: IEEE Computer Society Press, 2017: 1122-1131
- [40] Sun Y L, Lu A, Yu L. Weighted-to-spherically-uniform quality evaluation for omnidirectional video[J]. *IEEE Signal Processing Letters*, 2017, 24(9): 1408-1412
- [41] Zhou Y F, Yu M, Ma H L, *et al.* Weighted-to-spherically-uniform SSIM objective quality evaluation for panoramic video[C] //Proceedings of the 14th IEEE International Conference on Signal Processing. Los Alamitos: IEEE Computer Society Press, 2018: 54-57
- [42] Dong C, Loy C C, He K M, *et al.* Image super-resolution using deep convolutional networks[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, 38(2): 295-307
- [43] Dong C, Loy C C, Tang X O. Accelerating the super-resolution convolutional neural network[C] //Proceedings of the 14th European Conference on Computer Vision. Heidelberg: Springer, 2016: 391-407
- [44] Shi W Z, Caballero J, Huszar F, *et al.* Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 1874-1883
- [45] Zhang Y L, Tian Y P, Kong Y, *et al.* Residual dense network for image super-resolution[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 2472-2481
- [46] Ozcinar C, Rana A, Smolic A. Super-resolution of omnidirectional images using adversarial learning[C] //Proceedings of the 21st IEEE International Workshop on Multimedia Signal Processing. Los Alamitos: IEEE Computer Society Press, 2019: 1-6