引用格式：杜丽娜，杨硕，卓力，等. 基于 ResNet-TSM 和 BiGRU 网络的移动视频感知质量评价模型[J]. 北京工业大学学报，2024，50(1)：18-26.

DU L N，YANG S，ZHUO L，et al. Mobile video perceptual quality assessment model with ResNet-TSM and BiGRU network[J]. Journal of Beijing University of Technology，2024，50(1)：18-26.（in Chinese）

基于 ResNet-TSM 和 BiGRU 网络的移动视频感知质量评价模型

杜丽娜[1,2]，杨 硕[1,2]，卓 力[1,2]，张 菁[1,2]，李嘉锋[1,2]

（1. 北京工业大学信息学部，北京 100124；2. 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）

摘 要：考虑到卡顿、质量切换、内容特征等因素对用户体验质量的影响都会直接体现在客户端的失真视频里，提出了一种客户端的移动视频感知质量评价模型。该模型无须对每种影响因素均进行表征和度量，而是基于深度特征提取 + 回归的思路，直接建立失真视频与平均意见分数之间的映射模型。首先，构建了 ResNet-TSM 网络结构，提取失真视频片段的深度时空特征；为了避免维度灾难，采用 LargeVis 算法对提取的深度特征进行降维，同时提升特征的表达与区分能力。然后，采用双向门控循环单元网络对视频的长时间依赖关系进行建模，得到各视频片段的打分，再利用时间平均池化方法将各片段分数进行聚合，得到整个视频的打分结果。在 WaterlooSQoE-Ⅲ和 LIVE-NFLX-Ⅱ数据集上的实验结果表明，提出的模型可以获得更高的预测精度。

关键词：视频感知质量评价；平均意见分数；卷积神经网络；时间移位模块；双向门控循环单元；深度时空特征

中图分类号：TP 391 **文献标志码**：A **文章编号**：0254 - 0037(2024)01 - 0018 - 09

doi：10.11936/bjutxb2022020009

Mobile Video Perceptual Quality Assessment Model With ResNet-TSM and BiGRU Network

DU Lina[1,2]，YANG Shuo[1,2]，ZHUO Li[1,2]，ZHANG Jing[1,2]，LI Jiafeng[1,2]

（1. Faculty of Information Technology，Beijing University of Technology，Beijing 100124，China；
2. Beijing Key Laboratory of Computational Intelligence and Intelligent System，Beijing University of Technology，Beijing 100124，China）

Abstract：Considering the effects of stalling，quality switching，content characteristics and other factors，which will be directly reflected in the distorted video，a client-oriented mobile video perceptual quality assessment model was proposed. The mapping model between the distorted video and the mean opinion score（MOS）was established based on the idea of “deep feature extraction + regression” instead of characterizing and measuring each influencing factor. First，ResNet-TSM network was constructed to extract the deep spatial-temporal features of each distorted video segmentation. Second，LargeVis

收稿日期：2022-02-24；修回日期：2022-03-24

基金项目：国家自然科学基金资助项目(61531006)；北京市自然科学基金资助项目(KZ201910005007)

作者简介：杜丽娜(1995—)，女，博士研究生，主要从事视频质量评价、码率自适应算法方面的研究，E-mail：1340398672@qq.com

通信作者：卓 力(1971—)，女，教授，博士生导师，主要从事图像/视频的编码与传输、多媒体大数据处理方面的研究，E-mail：zhuoli@bjut.edu.cn

algorithm was used to reduce the dimensionality of the extracted deep features, and simultaneously improving the representation and discriminative capabilities of the features. Afterward, the score of each video segment was obtained by modeling the long-term dependence of the video by using the bidirectional gated recurrent unit. The temporal mean pooling was adopted to aggregate the scores of each segment to obtain the overall video score. The experimental results on the WaterlooSQoE-Ⅲ and LIVE-NFLX-Ⅱ datasets show that the proposed model can achieve a higher prediction accuracy.

Key words: video perception quality assessment; mean opinion score; convolutional neural network; time shift module; bidirectional gated recurrent unit; deep spatial-temporal features

近年来，随着移动网络和智能终端的发展，尤其是 5G 技术的出现，移动视频业务出现了迅猛的增长。思科视觉网络指数（visual networking index, VNI）[1]指出，到 2022 年全球移动数据流量的近 4/5 将来自视频业务，从 2017 年至 2022 年移动视频流量将增长 9 倍。对于移动视频业务供应商而言，有效利用可用资源进行视频流的传输，同时保证用户的体验质量（quality of experience, QoE）是在激烈的市场竞争中取得成功的关键。

目前移动视频流传输普遍采用基于 HTTP 的自适应流媒体协议（dynamic adaptive streaming over HTTP, DASH）[2]。根据 DASH 协议，在服务器端，视频被切分为多个视频片段，每个片段分别采用不同的码率进行编码并存储。在客户端，视频播放器可以根据网络可用带宽、缓冲区状态等因素自适应确定下一个要播放的视频片段的码率。但是当上述因素发生剧烈波动时，会导致视频质量频繁切换甚至会导致视频播放出现中断，从而严重影响用户的 QoE[3]。因此，如何对 QoE 进行评价打分，以准确反映出用户观看视频的主观体验，就受到工业界和学术界的广泛关注。

QoE 建模的实质在于找到一种映射关系 $Y=f(X)$，其中 X 是影响用户 QoE 的各种因素，Y 是用户的主观感受质量，通常用平均意见得分（mean opinion score, MOS）来度量。近年来，人们提出了各种 QoE 预测模型[4-9]，主要考虑的影响因素有视频编码码率、卡顿、质量切换、内容特性等，首先分别对每种影响因素进行量化和表征，然后采用各种机器学习的方法建立 Y 和 X 之间的关系模型。

考虑到视频编码码率、卡顿、质量切换、内容特征等对用户 QoE 的影响都直接体现在用户观看的失真视频里，本文提出了一种客户端的移动失真视频感知质量评价模型。该模型采用 MOS 对视频的感知质量进行度量，无须对各种影响因素分别进行度量和表征，而是基于“深度特征提取 + 回归”的思路，直接建立失真视频与 MOS 之间的映射模型，用于对失真视频的感知质量进行预测。

本文将提出的模型在 2 个公开的视频数据集上进行评估，并将其与其他最新的视频感知质量评价模型进行比较，对比结果可以验证本文方法的有效性。

1　相关工作

对于移动视频业务来说，影响用户 QoE 的因素有很多，既包含主观因素，如兴趣偏好、习惯等，又包含客观因素，如网络状况、卡顿、缓存状态等。这些影响因素相互作用，给 QoE 评价带来了极大的难度和挑战。

近年来，人们提出各种 QoE 评价模型，考虑的影响因素 X 主要有视频质量、卡顿、质量切换、内容特征等，然后采用各种机器学习的方法建立 Y 和 X 之间的关系模型。比如，文献[4-5]考虑码率、卡顿因素和质量切换因素，分别采用线性回归和对数回归的方法建立了模型；文献[6-7]采用全参考的结构相似指数（structural similarity index plus, SSIMplus）[10]度量视频质量，同时考虑了视频的卡顿因素，建立了线性回归模型；P. 1203[8]考虑视频码率、分辨率、卡顿和质量切换等因素，采用随机森林建立了评价模型；Duanmu 等[9]以全参考的视频多方法评估融合（video multimethod assessment fusion, VMAF）[11]度量视频的质量，同时考虑卡顿因素和质量切换，采用线性回归的方法建立了模型。

总地来看，现有的 QoE 评价模型往往需要分别提取每个影响因素的特征参数，用于对各种影响因素进行量化和表征，然后将多个特征参数组合后形成影响因素特征向量，并采用机器学习的方式建立特征向量和 MOS 之间的映射关系模型。由于各种影响因素之间存在复杂的关联关系，这种方式很难对影响因素进行有效的量化和表征。

就建模方法而言，需要强调的是用户在观看视

频时的记忆效应也会对感知质量的评估造成影响。Bampis 等[12]指出,首因效应和近因效应会在一定程度上影响用户的观看体验,也就是说在视频感知质量评价过程中还应该考虑视频流的长时间依赖关系。文献[13-14]已经证明了长短期记忆(long short-term memory, LSTM)[15]网络可以在连续性视频感知质量度量中获得更好的性能。由于 LSTM 的复杂结构会导致训练的困难,Cho 等[16]采用门控循环单元(gated recurrent unit, GRU)网络很好地解决了这一问题。

考虑到 QoE 的各种影响因素之间存在复杂的关联关系,而视频编码码率、卡顿、质量切换、内容特征等对用户主观感知质量的影响都直接体现在用户观看的失真视频里。因此,本文提出了一种客户端的移动视频感知质量评价模型。该模型无须对 QoE 的每种影响因素均进行表征和度量,而是直接建立失真视频与 MOS 之间的映射模型。实验结果表明,与现有的 QoE 评价模型相比,本文提出的模型无须其他的信息,还可以获得更好的性能。

2 失真视频感知质量评价模型

本文提出的失真视频感知质量评价模型整体框架如图 1 所示。为了降低存储和计算压力,将失真视频切分成等长的片段,对每个视频片段分别进行特征提取,然后对各个片段的深度特征进行回归,并进一步聚合得到整个视频的感知质量打分结果。

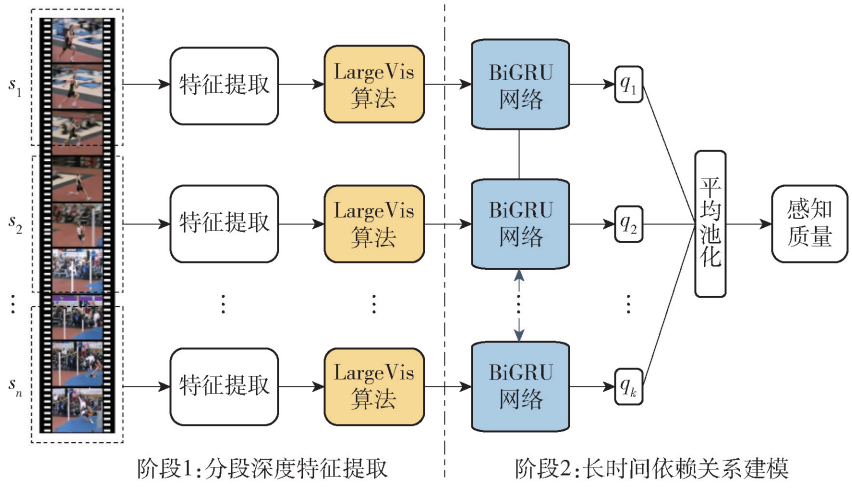

图 1 视频感知质量评价模型框架

Fig. 1 Framework of video perceptual quality assessment model

本文提出的模型主要包括 2 个核心部分:1)失真视频片段的深度特征提取。本文采用 ResNet-50[17]作为骨干网络,将时间移位模块(temporal shif module,TSM)[18]引入其中,构建了 ResNet-TSM 网络结构,用于提取视频片段的深度时空特征,然后采用 LargeVis 算法[19]对深度特征进行降维,减轻后续的计算和存储压力。2)对视频片段之间的长时依赖关系进行建模,建立视频片段的感知质量评价模型,用于对 MOS 进行预测。最后将每段视频的 MOS 打分结果进行时间平均池化,获得整个视频的感知质量评价打分结果。下面将分别介绍每个部分的实现细节。

2.1 基于 ResNet-TSM 网络的深度时空特征提取

2.1.1 深度时空特征提取

目前提取视频深度时空特征普遍采用三维卷积神经网络(three-dimensional convolutional neural network,3D CNN),其计算复杂度高,所需存储空间大。为了实现计算效率和性能之间的良好折中,本文构建了 ResNet-TSM 网络结构,用于提取视频的深度时空特征,深度时空特征提取框图见图 2,在 TSM 中,T 表示提取的特征图的时间维度,C 表示特征图的通道维度,H 和 W 分别表示特征图的高和宽。

TSM 核心思想是将特征图的部分通道沿时间维度进行前后的移位,因此在某一时刻的特征图中会包含相邻帧的信息,可以表达视频的时间特性。另外,由于通道的移动会损害视频帧的空间信息,为了保证空间信息的完整性,TSM 通常采用残差的结构插入二维卷积神经网络(two-dimensional convolutional neural network,2D CNN)中。本文采用 ResNet-50 作为骨干网络,在每个残差单元中都加入了 TSM,构建了 ResNet-TSM 网络结构。由于 ResNet50 提取的特征图只能表示空间特征,而 TSM

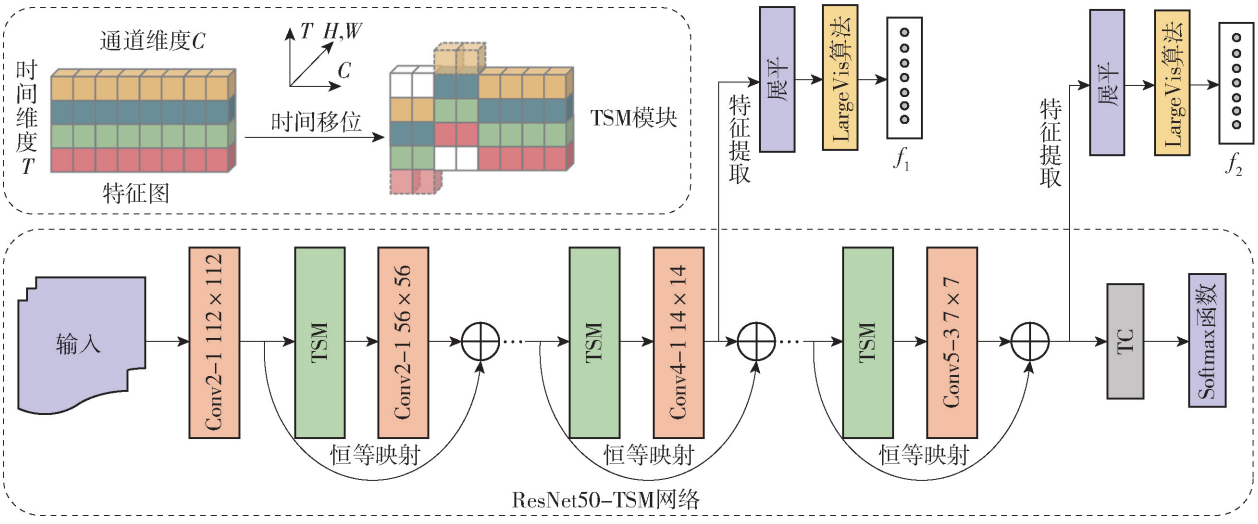

图 2　基于 ResNet-TSM 网络的深度特征提取框图

Fig. 2　Framework of deep feature extraction with ResNet-TSM

会将 t 时刻视频帧的特征图和 $t-1$、$t+1$ 时刻的特征图进行部分通道的移位操作，因此 t 时刻的特征图包含 $t-1$ 和 $t+1$ 时刻视频帧的信息。虽然移位操作会影响二维空间维度的特征信息，但是通过残差操作可以使原始的二维空间特征得到保留；由于移位运算基本不消耗计算资源，因此能够以 2D CNN 的复杂度实现 3D CNN 的性能。

为了提升网络训练性能，本文采用了预训练 + 微调的策略对 ResNet-TSM 网络进行训练。首先利用 ImageNet 数据集[20]对 ResNet 网络进行预训练，训练后的网络参数作为 ResNet-TSM 网络的初始模型参数。然后，采用失真视频数据集对初始模型参数进行微调，得到优化后的 ResNet-TSM 网络模型。

在训练过程中，本文根据失真视频片段的 VMAF 分数，将其分为 5 个等级：0 ~ 20 分为难以接受，20 ~ 40 分为较差，40 ~ 60 分为一般，60 ~ 80 分为较好，80 ~ 100 分为优秀[13]。属于一个等级的视频帧为一类，整个数据集共有 5 个类别，用于微调网络参数。在微调过程中，对于每个视频片段，从中随机抽取 8 帧，然后将各帧归一化为 224 × 224 大小，输入到网络中进行训练。

对于卷积神经网络而言，通常低卷积层可以提取边缘、纹理之类的低层视觉特征，因此本文首先提取了 ResNet-TSM 网络中 Conv4_1 层输出，用于表征视频的纹理特征，其维度为 8 × 14 × 14 × 256。由于每次插入 TSM 后网络的时间感受野都会扩大 2 个单位，故整个框架最后的时间感受野会很大。为了获取足够的时间信息，本文提取 ResNet-TSM 网络中全连接（fully connected，FC）层之前的 8 × 2 048 维输出，用于表征该视频片段的时间特征。2 个特征级联起来形成特征向量，用于表征视频片段的时空内容特性。

2.1.2　LargeVis 算法特征降维

从 ResNet-TSM 网络中提取的时空特征维度分别是 8 × 14 × 14 × 256 和 8 × 2 048，如此高的特征维度会造成维度灾难。为此，本文采用了 LargeVis 算法对其进行降维。LargeVis 算法是 Tang 等[19]提出的一种基于概率的降维算法，该算法是面向大数据可视化设计的，其基本思想是在高维空间中距离接近的数据点投影到低维空间后也要保持这种近邻关系，而数据点之间的距离通过条件概率来度量。该方法不仅能有效去除数据中的冗余信息，还可以有效地拉大原来在高维空间中相距较远的不同类别簇之间的距离，缩小类内样本之间的距离，能够有效提升特征的区分能力。

本文将提取的视频片段的时空特征维度首先从 8 × 14 × 14 × 256 和 8 × 2 048 展平为 401 408 和 16 384，然后利用 LargeVis 算法将两者的特征维度降至 128 维，获得一个紧凑的特征表达。最后，本文将降维后的特征级联起来，得到视频片段的深度时空特征表示向量 $\boldsymbol{F}$，可以表示为

$$\boldsymbol{F} = C(f_1, f_2) \tag{1}$$

式中：C 表示级联操作；f_1 和 f_2 分别表示降维后的视频片段的空间和时间特征；$\boldsymbol{F}$ 的维度是 256。

2.2　视频长时间依赖关系建模

用户在对视频质量进行评判时会受到时间记忆的影响，这意味着在对视频感知质量进行建模的过程中需要考虑复杂的时间依赖关系。Cho 等[16]提

出了一种 LSTM 网络的变体,称为 GRU 网络,其结构比标准 LSTM 网络的简单,可以在保证精度基本不变的同时减少参数量,加快训练速度。考虑到单向的 GRU 网络仅考虑前向依赖性,很可能会丢失或无法转发一些有用信息,因此本文采用了 BiGRU 网络来建模视频的长时依赖关系,其网络框架如图 3 所示。将视频片段的深度特征向量 $\boldsymbol{F}_k$ 送入 BiGRU 网络,并将 BiGRU 网络的输出作为集成特征。其中,当前 BiGRU 单元的输出 $\boldsymbol{Y}_k$ 的公式为

$$\boldsymbol{Y}_k = \mathrm{BiGRU}(\boldsymbol{Y}_{k-1}, \boldsymbol{F}_k, \boldsymbol{Y}_{k+1}) \tag{2}$$

式中:$\boldsymbol{F}_k$ 是当前视频片段的深度时空特征;$\boldsymbol{Y}_{k-1}$ 和 $\boldsymbol{Y}_{k+1}$ 分别表示前一个和后一个 BiGRU 单元的输出。本文中,单个 GRU 单元的输入和输出维度被分别设置为 256 和 32,即输入 $\boldsymbol{F}_k$ 为 256;输出 $\boldsymbol{Y}_k$ 为 64。然后将集成的特征 $\boldsymbol{Y}_k$ 送入 2 个 FC 层,进而得到每个视频片段的感知质量打分结果,其中,第 1 个 FC 层(FC1)中的神经元数为 32 个,第 2 个 FC 层(FC2)中的神经元数为 1 个。

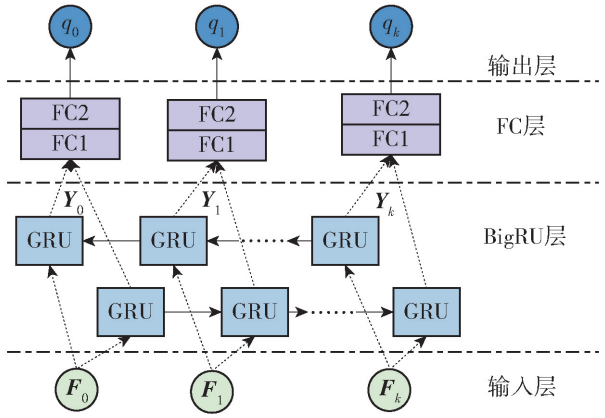

图 3 BiGRU 网络结构

Fig. 3 Architecture of BiGRU network

为了将视频片段的分数进行聚合,本文对各个片段的感知质量打分结果进行时间平均池化操作,得到视频整体的感知质量。时间平均池化可用公式表示为

$$Q = \frac{1}{N}\sum_{k=1}^{N} q_k \tag{3}$$

式中:N 表示视频片段的数目;q_k 表示第 k 个视频片段的 MOS。

3 实验结果与分析

为了验证所提出评价模型的有效性,本文在 WaterlooSQoE-Ⅲ[21] 和 LIVE-NFLX-Ⅱ[22] 2 个公开的视频数据集上进行了实验。下面介绍实验结果。

3.1 数据集和性能评价指标

WaterlooSQoE-Ⅲ数据集包含 20 个不同内容类型的原始高质量视频,采用 H.264 编码器对原始视频分别以 11 个固定码率等级(235 ~ 7 000 kbit/s)进行编码,之后将其存储于服务端,客户端根据 DASH 协议采用 6 种具有代表性的 ABR 算法在 13 种网络环境下进行仿真,获得了 450 个具有不同失真程度的视频,失真视频的平均持续时长为 13 s。最后,根据国际电信联盟主观打分标准,采用单激励方法,经由 34 个测试人员的打分获取了每个失真视频的 MOS。LIVE-NFLX-Ⅱ数据集则包含 15 个不同类型的原始视频,原始视频根据内容驱动的动态优化器进行码率编码(150 ~ 1 000 kbit/s),同时采用 4 种码率自适应算法在 7 种不同移动网络条件下进行仿真,生成了包含连续性和回顾性得分的 420 个失真视频,平均视频时长为 25 s。上述 2 个数据集均采用 MOS 作为整个视频感知质量的度量指标,其中 LIVE-NFLX-Ⅱ数据集还进一步提供了每个失真视频帧的 MOS。

为了评估算法的性能,本文采用了最常用的 3 个性能评估指标:皮尔森线性相关系数(Pearson's linear correlation coefficient, PLCC)、斯皮尔曼秩相关系数(Spearman rank-order correlation coefficient, SROCC)和肯德尔秩次相关系数(Kendall rank-order correlation coefficient, KROCC)。3 个指标的数值越高,则表示模型的预测性能越好。PLCC、SROCC 的定义分别为

$$V_{\mathrm{PLCC}} = \frac{\sum_{i=1}^{n}(y_{\mathrm{p}i}-\bar{y}_{\mathrm{p}})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(y_{\mathrm{p}i}-\bar{y}_{\mathrm{p}})^2(y_i-\bar{y})^2}} \tag{4}$$

$$V_{\mathrm{SROCC}} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{5}$$

式中:$y_{\mathrm{p}i}$ 和 y_i 分别表示第 i 个视频的预测分数和真实分数;$\bar{y}_{\mathrm{p}}$ 和 $\bar{y}$ 分别表示预测平均值和真实平均值;n 表示测试集中视频的个数;d_i 表示第 i 个视频的预测分数和真实分数之间的差异量。KROCC 的定义为

$$V_{\mathrm{KROCC}} = (S-R)/N \tag{6}$$

对于 KROCC,将预测分数 $y_{\mathrm{p}i}$ 和真实分数 y_i 组成 n 个数据对,从 n 个数据对中任取 2 个数据对 $\{(y_{\mathrm{p}i}, y_i), (y_{\mathrm{p}j}, y_j)\}$,$(i \neq j)$,共有 $N = n(n-1)/2$ 种可能。对任取的 2 个数据对,若 $y_{\mathrm{p}i} > y_i, y_{\mathrm{p}j} > y_j$ 或

者 $y_{pi}<y_i, y_{pj}<y_j$，则这样的数据对被称为同序数据对；若 $y_{pi}>y_i, y_{pj}<y_j$ 或者 $y_{pi}<y_i, y_{pj}>y_j$，则这样的数据对被称为逆序数据对；S 表示同序数据对的个数，R 表示逆序数据对的个数。

3.2　参数设置

本文将失真视频数据集随机切分为 2 个子集，其中 80% 用于训练，20% 用于测试。在分段视频特征提取阶段，本文将视频切分成长度为 3 s 的若干个视频片段。WaterlooSQoE-Ⅲ数据集中的失真视频被切分为 4 个片段，LIVE-NFLX-Ⅱ数据集的视频则被切分为 8 个片段。模型的训练分为两部分进行，其相关参数设置如下：

1）在特征提取阶段。利用切分后的失真视频片段对预训练后的 ResNet-TSM 网络进行微调。在选择特征通道进行移位的过程中，为了避免移位操作对空间特征造成过多的损伤，同时获取更多的时间特征，本文选择将特征图中 1/8 的通道进行前向移位，1/8 的通道进行后向移位，从而获取视频帧之间的前后向时间依赖关系。微调时训练时迭代次数设置为 70，初始学习率为 1×10^{-3}，批次大小为 12。

2）在视频的长时依赖关系建模阶段，本文选择将 LargeVis 算法降维后获得维数为 256 的深度时空特征作为每个 BiGRU 网络单元的输入，输出为 MOS。在训练过程中迭代次数、初始学习率和批次大小分别设置为 3 000、6×10^{-4} 和 16。

3.3　LargeVis 算法降维对模型精度的影响

如前所述，每个视频片段提取的特征维度分别为 $8\times14\times14\times256$ 和 $8\times2\,048$，这么高的维度会给后续的计算和存储带来极大的压力。为了避免维度灾难，本文采用 LargeVis 算法对视频的深度特征进行降维。为了研究不同特征维度对模型性能的影响，本文将每个特征的维度分别降维至 32、64、128、256 和 512，然后再进行建模，实验结果如表 1 所示。可以看出，当特征维度降至 128 时，所提出的模型在 2 个数据集上均能获得较好的性能。综合考虑模型复杂度和模型的性能，本文选择将 LargeVis 算法降维后的维度设置为 128。

表 1　LargeVis 算法降维对模型性能的影响

Table 1　Influence of LargeVis dimensionality reduction on QoE model accuracy

特征维度	WaterlooSQoE-Ⅲ			LIVE-NFLX-Ⅱ		
	V_{PLCC}	V_{SROCC}	V_{KROCC}	V_{PLCC}	V_{SROCC}	V_{KROCC}
32	0.892 3	0.852 5	0.669 9	0.948 4	0.939 3	0.791 2
64	0.877 9	0.850 1	0.674 4	0.943 1	0.937 6	0.795 2
128	0.906 0	0.862 2	0.700 4	0.958 3	0.955 9	0.819 3
256	0.906 6	0.852 3	0.679 4	0.957 7	0.947 3	0.806 4
512	0.901 2	0.861 5	0.691 7	0.960 0	0.947 8	0.815 3
未降维	0.869 1	0.833 4	0.674 1	0.912 2	0.903 7	0.789 6

3.4　不同网络提取的特征参数对模型性能的影响

为了进一步研究 ResNet-TSM 网络和 ResNet 网络提取的特征对于模型性能的影响，本文提取了 ResNet-TSM 网络和 ResNet 网络中 Conv2_1 到 Conv4_1 的特征，并采用 LargeViS 算法将维度降到 128，分别建立了视频感知质量评价模型。在 2 个数据集上的实验结果如表 2 所示。对比表中 2 个网络 Conv2_1、Conv3_1 与 Conv4_1 的结果可以看出，提取 Conv4_1 特征建立的模型性能最优。同时，提取 2 个网络低层特征建立的模型性能相差较小，这是由于 ResNet-TSM 网络主要增加了模型对于时间特性的表达能力，而空间表达能力则与 ResNet 网络相近。

3.5　不同特征参数对模型性能的影响

本文分别提取了 ResNet-TSM 网络的低层输出 f_1 和高层输出 f_2，分别表示视频片段的空间纹理特征和短时时间特征。为了研究 f_1 和 f_2 对模型性能的影响，本文对其进行了组合，分别建立了视频感知质量评价模型。在 2 个数据集上的实验结果如表 3 所示。

通过表 3 的结果可以看出：只考虑低层空间纹理特征 f_1 的情况下，模型准确度最低，这是由于缺乏视频片段的时间信息。相比而言，深层的特征 f_2 可以很好地表征视频片段的时间信息，与 f_1 相比，在 WaterlooSQoE-Ⅲ数据集上 V_{PLCC}、V_{SROCC} 和 V_{KROCC}

表 2 不同特征参数对模型性能的影响

Table 2 Influence of the combination of different influencing factors on model performance

网络结构	特征参数	WaterlooSQoE-Ⅲ			LIVE-NFLX-Ⅱ		
		V_{PLCC}	V_{SROCC}	V_{KROCC}	V_{PLCC}	V_{SROCC}	V_{KROCC}
ResNet-TSM	Conv2_1	0.681 3	0.665 1	0.483 0	0.798 1	0.783 6	0.562 6
	Conv3_1	0.714 2	0.687 3	0.507 2	0.820 5	0.812 1	0.617 8
	Conv4_1	0.750 5	0.722 2	0.540 1	0.868 0	0.856 7	0.678 1
ResNet50	Conv2_1	0.663 8	0.644 1	0.462 2	0.770 4	0.760 6	0.546 3
	Conv3_1	0.678 5	0.668 9	0.479 4	0.788 1	0.788 1	0.563 1
	Conv4_1	0.691 1	0.687 6	0.497 1	0.795 2	0.783 3	0.570 4

表 3 不同影响因素组合对模型性能的影响

Table 3 Influence of the combination of different influencing factors on model performance

特征参数	WaterlooSQoE-Ⅲ			LIVE-NFLX-Ⅱ		
	V_{PLCC}	V_{SROCC}	V_{KROCC}	V_{PLCC}	V_{SROCC}	V_{KROCC}
f_1	0.750 5	0.722 2	0.540 1	0.868 0	0.856 7	0.678 1
f_2	0.874 4	0.817 7	0.650 9	0.948 2	0.937 4	0.794 6
$f_1 + f_2$	0.906 0	0.862 2	0.700 4	0.958 3	0.955 9	0.819 3

分别提升 12.39%、9.55% 和 11.08%，在 LIVE-NFLX-Ⅱ数据集上 3 个指标分别提升 8.02%、8.07% 和 11.65%。而 f_1 和 f_2 组合起来可以很好地表征视频片段的时空内容特征，与仅利用 f_1 或者 f_2 相比，模型可以获得最优的预测精度。

3.6 不同建模方法的性能对比

为了验证不同建模方法对于模型性能的影响，本文以组合后的深度时空特征向量作为输入，分别采用决策树、支持向量回归（support vector regression，SVR）等 8 种浅层机器学习的方法和 4 种可以建模长时间依赖关系的方法来建立感知质量预测模型。对比实验结果如表 4 所示。从表 4 的结果可以看出：

1）浅层机器学习方法中，随机森林方法可以获得最优的性能。

2）相比于浅层机器学习方法，LSTM 等 4 种对长时依赖关系进行建模的网络可以获得更优的性能。

3）相比于 LSTM 网络，采用 GRU 进行建模，模型的预测准确度更高，在 WaterlooSQoE-Ⅲ数据集上 V_{PLCC}、V_{SROCC} 和 V_{KROCC} 分别提升了 0.45%、0.74% 和 0.75%，在 LIVE-NFLX-Ⅱ数据集上 3 个评估指标分别提升 0.85%、1.11% 和 2.19%。

4）通过分别对比 LSTM、BiLSTM 网络，GRU、BiGRU 网络可以看出，相比于单向网络，双向网络的表现更加出色。这是由于单向网络仅考虑前向依赖，可能会丢失或者无法转发一些重要的信息。但是需要指出的是，双向网络的结构复杂度、参数规模也要明显高于单向网络。

3.7 与不同 QoE 模型的性能比较

为了评估本文提出的模型性能，本文将其在 WaterlooSQoE-Ⅲ和 LIVE-NFLX-Ⅱ数据集上与现有的各种 QoE 评价模型进行了比较。在实验过程中，本文采用 LargVis 算法将视频片段的深度特征 f_1、f_2 的维度均降至 128，级联后深度特征的维度为 256，并采用 BiGRU 网络进行长时依赖关系的建模。参与对比的模型有 7 种，分别是：

1）Yin 2015[4] 和 Spiteri 2016[5]。采用编码码率作为视频质量表示，同时提取了卡顿和质量切换的特征参数，最后分别采用线性回归和对数回归的方法建立了预测模型。

2）Bentaleb 2016[6] 和 SQI[7]。以全参考 SSIMplus 度量视频质量，同时考虑了视频的卡顿因素，建立了线性回归模型。

3）P.1203[8]。提取了编码码率、分辨率、卡顿因素和质量切换因素等方面的特征形成影响因素向量，并采用随机森林的方法建立了预测模型。

4）KSQI[9]。以全参考 VMAF 指标来度量视频质量，结合卡顿因素和质量切换等方面的特征参数，采用线性回归方法建立了预测模型。

5）CGNN[23]。提取了图像空间质量、卡顿时长、帧率、质量切换等多种影响因素，采用 CNN 和 GRU 网络构建了深度神经网络，建立了预测模型。参与对比的模型在 2 个基准数据集上预测结果均引用自文献[9,23]，各个模型的对比结果如表 5 所示。可以看到：

表 4　采用不同建模方法得到的模型性能比较

Table 4　Comparison results of the performance with different modeling methods

建模方法	WaterlooSQoE-Ⅲ			LIVE-NFLX-Ⅱ		
	V_{PLCC}	V_{SROCC}	V_{KROCC}	V_{PLCC}	V_{SROCC}	V_{KROCC}
Decision Tree	0.681 2	0.650 2	0.484 8	0.891 2	0.885 3	0.708 2
Bagging Regressor	0.862 3	0.800 0	0.626 5	0.917 1	0.897 0	0.736 7
Bayesian Regressor	0.876 6	0.813 7	0.634 0	0.901 5	0.894 3	0.718 6
KNeighbors Regressor	0.825 7	0.747 8	0.604 0	0.884 9	0.873 9	0.681 4
SVR	0.869 0	0.816 8	0.644 4	0.902 1	0.886 2	0.706 2
Ridge Regressor	0.889 1	0.820 6	0.644 9	0.919 5	0.907 5	0.742 4
LassoCV	0.876 3	0.809 4	0.625 5	0.905 8	0.899 2	0.731 1
Random Forest	0.872 6	0.829 0	0.658 4	0.920 0	0.907 8	0.752 5
LSTM	0.903 1	0.837 9	0.672 4	0.948 7	0.939 0	0.787 1
BiLSTM	0.906 2	0.841 4	0.677 9	0.955 1	0.946 2	0.807 8
GRU	0.907 6	0.845 3	0.679 9	0.957 2	0.950 1	0.809 0
BiGRU	0.906 0	0.862 2	0.700 4	0.958 3	0.955 9	0.819 3

表 5　不同 QoE 模型的性能比较

Table 5　Modeling methods performance comparison with other state-of-the-art QoE models

QoE 模型	WaterlooSQoE-Ⅲ			LIVE-NFLX-Ⅱ		
	V_{PLCC}	V_{SROCC}	V_{KROCC}	V_{PLCC}	V_{SROCC}	V_{KROCC}
Yin2015[4]	0.722	0.714	0.543	0.673	0.686	0.482
Spiteri2016[5]	0.809	0.798	0.597	0.731	0.711	0.712
Bentaleb2016[6]	0.625	0.718	0.521	0.898	0.883	0.712
SQI[7]	0.673	0.690	0.496	0.910	0.906	0.735
P.1203[8]	0.769	0.797	0.604	0.817	0.821	0.619
KSQI[9]	0.794	0.776	0.584	0.905	0.893	0.722
CGNN-QoE[23]	0.890	0.881	0.707	0.935	0.927	0.778
本文方法	0.906 0	0.862 2	0.700 4	0.958 3	0.955 9	0.819 3

1）相比于其他模型，本文提出的模型获得了性能上的大幅提升，这是由于本文的方法提取的深度时空特征可以更有效地表示视频的内容特性，对于视频质量、卡顿、质量切换等因素对视频失真造成的影响可以很好地进行表征，同时也在一定程度上缓解了分别提取不同影响因素特征参数时难以有效表示不同特征之间的复杂关系这一问题。另外，本文采用的 BiGRU 网络非常适合从时变的视频中学习到长时依赖关系信息，从而获得更高的模型精度。

2）各个模型在 WaterlooSQoE-Ⅲ数据集上的预测准确度普遍低于 LIVE-NFLX-Ⅱ数据集，这与 WaterlooSQoE-Ⅲ数据集的失真情况比较复杂有关。

4　结论

1）采用 ResNet-TSM 网络提取的深度时空特征可以有效地表示视频的内容特性。

2）LargeVis 算法可以有效地对高维特征进行降维，在避免维度灾难的同时有效提升了特征的表达与区分能力。

3）BiGRU 网络具有同时学习视频的前向依赖和后向依赖关系的能力，可以有效表征视频片段间的长时间依赖关系，同时结构更加简单、参数更少且收敛性更好。

4）相比于其他模型，本文提出的模型获得了性能上的大幅提升。

参考文献：

[1] FORECAST G. Cisco visual networking index: global mobile data traffic forecast update, 2017—2022[R/OL]. [2022-02-22]. https://branden.biz/wp-content/uploads/2019/05/white-paper-c11-738429.pdf.

[2] MPEG I. Information technology-dynamic adaptive streaming over http (dash)-part 1: media presentation description and segment formats. [R/OL]. [2022-02-22]. https://www.w3.org/2010/11/web-and-tv/papers/webtv2_submission_64.pdf.

[3] AKHSHABI S, BEGEN A C, DOVROLIS C. An experimental evaluation of rate-adaptation algorithms in adaptive streaming over HTTP[C]// Proceedings of the Second Annual ACM Conference on Multimedia Systems. New York: Association for Computing Machinery, 2011: 157-168.

[4] YIN X, JINDAL A, SEKAR V, et al. A control-theoretic approach for dynamic adaptive video streaming over HTTP[C]// Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication. New York: Association for Computing Machinery, 2015: 325-338.

[5] SPITERI K, URGAONKAR R, SITARAMAN R K. BOLA: Near-optimal bitrate adaptation for online videos[J]. IEEE/ACM Transactions on Networking, 2020, 28(4): 1698-1711.

[6] BENTALEB A, BEGEN A C, ZIMMERMANN R. SDNDASH: improving QoE of HTTP adaptive streaming using software defined networking[C]// Proceedings of the 24th ACM international conference on Multimedia. New York: Association for Computing Machinery, 2016: 1296-1305.

[7] DUANMU Z, ZENG K, MA K, et al. A quality-of-experience index for streaming video[J]. IEEE Journal of Selected Topics in Signal Processing, 2016, 11(1): 154-166.

[8] RECOMMENDATION I. 1203. 3, Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport-quality integration module[R/OL]. [2022-02-22]. https://www.itu.int/rec/T-REC-P.1203.

[9] DUANMU Z, LIU W, CHEN D, et al. A knowledge-driven quality-of-experience model for adaptive streaming videos[J/OL]. arXiv preprint arXiv:1911.07944, 2019: 1-12[2022-02-22]. https://arxiv.org/abs/1911.07944.

[10] REHMAN A, ZENG K, WANG Z. Display device-adapted video quality-of-experience assessment[EB/OL]. Human Vision and Electronic Imaging XX, 2015, 9394: 939406[2022-02-22]. https://cin.ufpe.br/~cabm/visao/artigos/939406.pdf.

[11] LI Z, AARON A, KATSAVOUNIDIS I, et al. Toward a practical perceptual video quality metric[R/OL]. [2022-02-22]. https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652

[12] BAMPIS C G, LI Z, MOORTHY A K, et al. Study of temporal effects on subjective video quality of experience[J]. IEEE Transactions on Image Processing, 2017, 26(11): 5217-5231.

[13] ESWARA N, ASHIQUE S, PANCHBHAI A, et al. Streaming video QoE modeling and prediction: a long short-term memory approach[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 30(3): 661-673.

[14] DUC T N, TRAN C M, TAN P X, et al. Bidirectional LSTM for continuously predicting QoE in HTTP adaptive streaming[C]// Proceedings of the 2019 2nd International Conference on Information Science and Systems. New York: Association for Computing Machinery, 2019: 156-160.

[15] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[16] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, Penssylvania, USA: Association for Computational Liguistics, 2014: 1724-1734.

[17] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2016: 770-778.

[18] LIN J, GAN C, HAN S. TSM: Temporal shift module for efficient video understanding[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ, USA: IEEE, 2019: 7083-7093.

[19] TANG J, LIU J, ZHANG M, et al. Visualizing large-scale and high-dimensional data[C]// Proceedings of the 25th International Conference on World Wide Web. Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2016: 287-297.

[20] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]// 2009 IEEE conference on computer vision and pattern recognition. Piscataway, NJ, USA: IEEE, 2009: 248-255.

[21] DUANMU Z, REHMAN A, WANG Z. A quality-of-experience database for adaptive video streaming[J]. IEEE Transactions on Broadcasting, 2018, 64(2): 474-487.

[22] BAMPIS C G, LI Z, KATSAVOUNIDIS I, et al. Towards perceptually optimized adaptive video streaming-a realistic quality of experience database[J]. IEEE Transactions on Image Processing, 2021, 30: 5182-5197.

[23] ZHOU Z M, DONG Y, SONG L, et al. Quality of experience evaluation for streaming video using CGNN[C]// 2020 IEEE International Conference on Visual Communications and Image Processing. Piscataway, NJ, USA: IEEE, 2020: 285-288.

(责任编辑 郑筱梅)