视频质量评价研究综述

鄢杰斌 方玉明 刘学林 姚怡茹 眭相杰
（江西财经大学信息管理学院 南昌 330013）

摘 要 移动互联网时代每天都产生海量的质量参差不齐的视频数据，根据视频质量高效地过滤低质量视频对缓解设备存储压力起着至关重要的作用.此外，在视频的生成、处理、传输等过程中都不可避免地引入信号噪声，如何准确地预测视频质量，从而指导与监督视频处理与传输系统的优化具有重要的研究意义和实际价值.因此，视频质量评价受到越来越多的关注.视频质量评价旨在定量描述视频的视觉质量，包括主观质量评价和客观质量评价.主观质量评价通过开展视觉感知主观实验，研究各项因素对视觉质量的影响，并收集主观质量分数用于构建基准数据集；客观质量评价通过设计客观算法，自动预测视频的质量.本文首先介绍视频质量评价的基础知识，阐述视频质量评价的相关应用和问题；其次，重点介绍视频质量评价近二十年的发展现状，对比不同主观数据集的特点；然后，深入解析客观模型的建模思想，分层次对比不同的模型，详细分析各模型的优缺点；最后，指出未来发展方向并总结全文.

关键词 视频质量评价；视觉感知；特征工程；机器学习；深度学习
中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2023.02196

A Survey on Recent Advances in Video Quality Assessment

YAN Jie-Bin FANG Yu-Ming LIU Xue-Lin YAO Yi-Ru SUI Xiang-Jie
(*School of Information Technology, Jiangxi University of Finance and Economics, Nanchang* 330013)

Abstract In the current era of mobile internet, massive videos with varied quality have been produced daily. Therefore, it is of great importance to screen out low-quality videos quickly according to the predicted video quality to effectively relieve the storage pressure. In addition, distortion may be introduced inevitably in the procedure of video production, processing, transmission, display, and etc. Thus, estimating video quality accurately can be used for system optimization and algorithm optimization. Due to the above applications, video quality assessment (VQA) has gained more and more attention from both academia and industry. VQA aims to describe the quality of videos quantitatively, and it includes subjective quality assessment and objective quality assessment. The former means conducting psychophysical experiments, by which we can deeply explore the influence of different variables on video quality and collect subjective ratings for building benchmarking datasets, and the qualitative results of psychophysical experiments are often regarded as the guidance of designing objective VQA models. There are many mature and commonly used standards regarding collecting subjective ratings, such as single stimulus continuous quality evaluation and pair comparison, followed by outlier removal and final

收稿日期：2023-02-08；在线发布日期：2023-07-07. 本课题得到国家自然科学基金(No. 62132006)、中国博士后科学基金面上项目(No. 2022M721417)、江西省自然科学基金青年项目(No. 20224BAB212012)、江西省教育厅科技项目一般项目(GJJ2200522)资助.
鄢杰斌，博士，讲师，中国计算机学会(CCF)会员，主要研究领域为视觉质量评价、计算机视觉. E-mail：jiebinyan@foxmail.com. **方玉明**(通信作者)，博士，教授，中国计算机学会(CCF)会员，主要研究领域为多媒体信号处理、视觉质量评价. E-mail：leo.fangyuming@foxmail.com. **刘学林**，博士研究生，中国计算机学会(CCF)学生会员，主要研究领域为视觉质量评价. **姚怡茹**，硕士研究生，中国计算机学会(CCF)学生会员，主要研究领域为视觉质量评价. **眭相杰**，博士研究生，中国计算机学会(CCF)学生会员，主要研究领域为视觉质量评价.

quality score collection. The latter means designing objective VQA models for automatically and accurately predicting the quality of videos. According to the accessibility of reference information, VQA models could be classified into three categories, including full-reference (FR), reduced-reference (RR), and no-reference (NR). FR- and RR-VQA models need complete and partial reference information respectively when being deployed, and they often obey the 'similarity' measurement paradigm, i. e., the video which is more similar to the associated reference video is regarded to be of better visual quality. By contrast, NR-VQA models can predict the quality of videos without access to reference information, and they usually follow the 'feature extraction and quality regression' paradigm, where feature extraction may rely on priori knowledge (refers to hand-crafted features) or end-to-end deep learning technique (refers to deep semantic features) and quality regression may use shallow machine learning algorithm or deep neural network. According to the design philosophy, VQA models can be classified into three categories: natural scene statistics (NSS)-based, visual perception-based, and learning-based. NSS refers to the statistical regularities of visual scenes, where the statistical discrepancy between high-quality and test videos indicates the visual quality of the test video. The visual perception-based VQA models commonly simulate the complex perception process (e. g. the masking effect) of the human visual system by designing computational models. Unlike other two types of models, learning-based models usually construct the mapping function in a data-driven manner, i. e., using deep learning technique. In this paper, we first introduce the basic knowledge about VQA and describe the relevant applications and problems. Then, we focus on describing the development status of VQA in the past two decades, including pointing out the characteristics of different subjective databases, deeply analyzing and comparing the design philosophies of state-of-the-art VQA models, and introducing the pros and cons of these VQA models. Finally, we point out the potential development directions in the future and summarize this paper.

Keywords video quality assessment; visual perception; feature engineering; machine learning; deep learning

1 引 言

在当前的移动互联网时代，视频图像已经成为人们日常生产、生活中重要的数据来源[1-4]. 在视频的获取、传输、处理、存储、显示等过程中都可能出现扰动，导致视频质量出现下降，进而影响用户的视觉体验. 准确地度量视频的质量已经成为多媒体处理中重要的一项技术，受到工业界和学术界的广泛关注. 视频质量评价(Video Quality Assessment, VQA)指的是对视频信号进行分析，定量描述视频的视觉失真情况[5-6]. VQA包括主观质量评价和客观质量评价. 主观质量评价指的是开展大规模主观实验，研究各种影响因素对主观感知的影响及作用，为客观模型的设计提供理论基础，并为客观模型的性能计算提供基准；客观质量评价指的是构建VQA数学模型，能够自动地预测视频的质量，并期望获得与主观感知一致的预测结果. 根据参考信息依赖程度，VQA模型可分为：(1)全参考(Full Reference, FR)；(2)半参考(Reduced Reference, RR)；(3)无参考(No Reference, NR). 其中，FR-VQA模型和RR-VQA模型在计算视频质量时分别需要全部和部分参考信息，而NR-VQA模型在计算时不需要任何参考信息. 上述分类方法是根据视频参考信息的依赖程度对VQA模型进行分类，不涉及到VQA模型构建时依赖的知识. 另一种分类方法[7]是根据模型构建时依赖的先验知识来划分，将VQA模型大致分为三种：(1)依赖信号源先验知识的模型；(2)依赖人类视觉系统(Human Visual System, HVS)先验知识的模型；(3)依赖失真先验知识的模型. 其中，信号源知识指的是无失真信号视觉内容的本质，可直接从信号源或统计特性获取；HVS先验知识来源于视觉生理学和心理物理学研究；失真先验知识指的是失真类型及其特性，研究者可以根据失真类型及其

特性针对性地设计数学模型.VQA作为视频处理和视频理解领域的一个基础问题,它的具体应用[5-8]主要包括三点:

(1)数据筛选:比如在视频图像采集系统中,VQA模型可用于采集系统的视频图像质量监控;视频服务商可以根据视频的质量对视频进行筛选,去除质量较差的视频.另外,视频质量可以作为视频的重要属性用于其他任务,如视频检索等.

(2)参数选择.常用的参数选择是网格搜索(Grid Search, GS),即在候选的参数组合中寻找到最优的参数组合.在GS过程中,可以使用VQA模型预测的质量作为选择依据[9]-[10].当视频处理过程由多个算法组成时,在迭代处理的过程中可根据视频质量选择合适的算法序列,使得最终的处理结果视觉质量最好[11].另外,VQA算法可作为模型设计和验证的性能指标,用于比较不同的视频处理算法,从而确定性能最好的视频处理算法[6].

(3)模型/系统优化.在VQA研究不断发展的过程中,视频处理模型/系统优化获得了越来越多的关注.在模型优化中,最常用的VQA方法是均方误差(Mean Square Error, MSE)和结构相似性(Structural Similarity, SSIM)算法[12].针对不同的任务,可以使用合适的VQA算法作为损失函数去优化模型[13]-[17].另外,VQA算法可用于视频传输系统每个阶段的检测、优化和管理[18].

从早期研究中的合成失真、算法相关失真到真实失真,从视觉感知启发的VQA模型到数据驱动的深度VQA模型,VQA的发展呈现通用化和智能化.并且,随着主观数据集的丰富程度提高,VQA模型的性能也有显著的提升.然而,VQA的研究依然存在一些问题:

(1)数据集规模问题.相对于图像质量评价(Image Quality Assessment, IQA)[19-21]和其他图像处理和计算机视觉任务如视频图像分类[22-23]、视频理解[24]等,公开的视频质量主观数据集规模依然十分有限.据我们所知,最大的视频质量主观数据集包含不到4万个视频(含完整的主观标注),大部分视频质量主观数据集包含少于1千个视频.

(2)模型构建问题.相对于IQA模型,VQA模型在度量视频质量时一般会考虑时空信息或时域信息,常用的计算方法包括帧差法、光流法和卷积神经网络(Convolutional Neural Network, CNN)模型(如3D-CNN)等,往往忽略了空域信息、时空信息和时域信息在捕捉视频降质的作用[25].

(3)性能和效率兼顾问题.该问题是视频理解相关领域的一个经典问题,研究者针对这一问题提出了许多解决方案,包括针对输入[26-27]和模型[28-29]的设计.然而,VQA研究很少考虑该问题.

(4)应用问题.得益于多数IQA模型的易操作性、可导性以及IQA问题本身相对简易,IQA模型被广泛应用于其他领域,如多曝光图像融合、高动态范围图像色调映射、超分辨率、去噪、修复等[17].VQA在视频处理相关领域的应用相对局限,如视频编码[30-31]、传输[32]和增强[33-35]等.大部分应用仍然使用IQA模型作为客观评价指标[36-38].

研究者近二十年开展了大量的IQA和VQA研究工作,然而目前大部分相关综述论文都是针对IQA[39-44]的.虽然有少数针对VQA的综述[45-47],但是它们介绍的内容涵盖面较小.具体而言,文献[45]分别对IQA和VQA研究发展进行了描述,然而该文献仅介绍了针对二维视频[1]设计的VQA模型,忽略了很多最新的VQA模型,并且内容的区分也不够清晰;文献[46]详细介绍了主观质量评价和客观质量评价,但是该文献发表于2011年,未介绍最新的VQA研究进展;文献[47]是目前为止最新、最全面的VQA研究综述,但是它没有关于主观评价内容的介绍,且仅介绍了针对二维视频的研究工作而未介绍最新的VQA模型.考虑到现有的VQA综述论文[45-47]存在的问题并受到其他综述论文[48-51]的启发,本文系统性地介绍了VQA研究发展和最新进展,涵盖主观评价和客观评价.对于主观评价,本文梳理了测试数据的选择、主观数据收集方法及对比和各种内容视频质量主观数据集对比;对于客观评价,为了兼顾篇幅的平衡,本文根据视频内容的不同分别介绍二维视频和其他内容视频包括屏幕内容视频、三维视频、合成视频和全景视频客观模型(分别在第3节和第4节介绍);并从建模方式的角度对主流的VQA模型分门别类,详细介绍它们的建模思想和区别.本文内容框架如图1所示.

总体而言,本文的贡献在于:

(1)详细地介绍了主观质量评价方案,以及它们各自的适用场景;从建模的角度介绍了当前VQA模型的设计思路,并介绍了主流的和新近提出的评价指标.

(2)完整地介绍了现有的视频质量主观数据集,并对比了不同数据集的构建策略及特点.

1 指普通2D平面视频,用于区分其他内容视频.

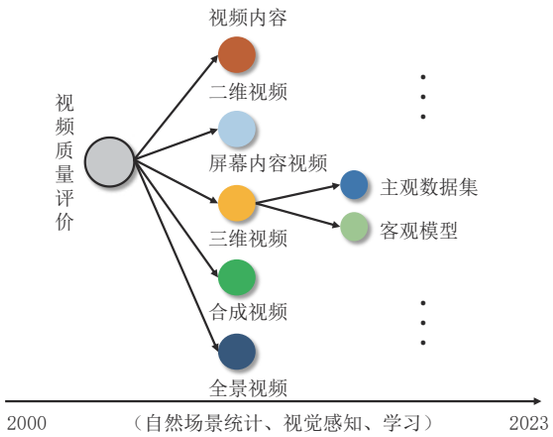


图1　本文内容框架

（3）系统地梳理了针对各类视频设计的客观质量评价模型，深入解析各个模型的设计原理及思路，并指出当前发展现状和未来发展趋势.

本文的后续章节的内容安排如下：第2节介绍主观实验、客观模型和评价指标基础知识；第3节详细介绍二维视频质量评价，描述当前发展现状和客观模型的设计理念；第4节详细介绍其他类型视频质量评价，重点突出各类视频的特点和相应的客观模型设计思想；第5节指出未来潜在的研究方向；第6节对全文进行总结.

2　基础知识概述

2.1　主观质量评价

不同于其他任务如分类、分割和检测等的标注过程，视频图像主观质量评价的不确定性较高，具体表现在：受试者对质量较好/差的视觉信号的评价一致性较高，而对于质量一般的视觉信号的评价一致性相对较低[52]. 因此，在视频图像质量评价领域中主观质量评价也是一个重要的研究课题. 一般而言，主观质量评价包括多个步骤：（1）测试数据的选择；（2）实验环境的选择；（3）主观数据的收集. 本文重点介绍第1点和第3点，原因在于当前构建大规模数据集的方法是众包[53-54]，即将主观标注任务以外包的形式分配给大众志愿者，而每个志愿者进行主观实验时的环境差异很大. 相对于在标准的实验环境下开展主观实验，众包方式的优势在于能够开展大规模主观实验，并且可以加快标注进程；它的不足在于主观数据的处理更加困难，因为大众志愿者的可信度更低[55-56]. 标准的实验环境设置可参考国际电信联盟的建议[57-58].

对于测试数据的选择，准则[57-58]包括：（1）数据的多样性，即空域感知信息（Spatial Perceptual Information，SI）和时域感知信息（Temporal Perceptual Information，TI）应该涵盖所有的范围[59]；（2）为了避免给受试者带来疲惫感从而保证主观数据的可靠性，测试数据至少应该包含4种不同的场景. SI和TI两种信息度量方式因计算简单，被广泛应用在VQA研究中，它们的计算方式如下：

$$SI = \max_{time} \{ std_{space}[\mathrm{Sobel}(F_n)] \} \tag{1}$$

$$TI = \max_{time} \{ std_{space}[M_n(i,j)] \} \tag{2}$$

$$M_n(i,j) = F_n(i,j) - F_{n-1}(i,j) \tag{3}$$

其中，F_n表示第n个时刻的视频帧；$Sobel(\cdot)$表示Sobel边缘检测算子；std表示标准差；space和time分别表示空域和时域. 除了SI和TI，研究者也有使用其他信息来保证数据的多样性，包括传统手工特征[60]如亮度、颜色、对比度、时域高斯导数、模糊等和深度特征[61-62]以及内容特征[63]等.

对于主观数据的收集，主观分数包括平均主观得分（Mean Opinion Score，MOS）和平均主观得分差（Differential Mean Opinion Score，DMOS），MOS的值越大表示质量越高，DMOS的值越小表示质量越高. 主观质量评价方法有多种，常见的有：双刺激连续质量量表（Double Stimulus Continuous Quality Scale，DSCQS）方法、双刺激损伤量表（Double Stimulus Impairment Scale，DSIS）方法、单刺激连续质量评测（Single Stimulus Continuous Quality Evaluation，SSCQE）方法和配对比较（Pair Comparison，PC）方法[57-58]. 它们的介绍具体如下：

（1）DSCQS方法给受试者呈现原始参考视频/图像和失真视频/图像对，其中原始参考视频/图像和失真视频/图像出现的顺序是随机的，受试者对原始参考视频/图像和失真视频/图像都进行打分.

（2）DSIS方法也称为损伤种类评分（Degradation Category Rating，DCR）. 该方法先后给受试者呈现原始参考和失真视频/图像对，受试者根据原始参考视频/图像对失真视频/图像打分.

（3）SSCQE方法也称为绝对种类评分（Absolute Category Rating，ACR）. 该方法仅给受试者呈现一次测试数据，且测试数据随机出现. 每次观看结束后，受试者给测试数据打分.

（4）PC方法中的测试数据成对出现，受试者需要指出成对测试数据的相对好坏. 相对于其他主观质量评价方法，该方法得到的主观判断一致性更高，

且打分速度更快. 但是该方法的主观判断的细粒度更低，因为它忽略了测试数据相对好坏的程度. 它的另一个缺点是当测试数据增加时成对比较的次数将呈爆炸式增长.

除了上述的主观数据收集方法，2018年Li等人[64]提出一种新颖的主观质量评价方法，称为AccAnn. 该方法用于收集体验质量(Quality of Experience，QoE)分数. 它是一个包含单一步骤的主观质量评价方法，质量等级设置为3. 相对于多步骤打分方法，它的优点在于：(1)打分速度快；(2)更容易理解. 该方法也被Yan等人[25]应用在自由视点视频(Free Viewpoint Video，FVV[1])QoE主观实验中. 主观评价方法的比较如表1所示.

表1 主观评价方法对比

名称	需要参考源	适用场景[58]	应用频率
DCSQ	√	测量系统相对某一基准的质量	中
DSIS	√	测量系统的牢靠程度	中
SSCQE	×	基准源信号不存在	高
PC	×	评价算法相对好坏	中

2.2 客观质量评价

本文在后续章节介绍VQA客观模型时参照该分类方法[7]，即从构建模型时依赖的先验知识角度对经典的VQA模型和当前主流的VQA模型分门别类，并进行详细地介绍. 考虑到大部分VQA模型都是通用型而非针对某种失真设计的，且基于学习(主要是深度学习)的VQA模型受到了科研人员越来越多的关注，本文将VQA模型分成三类：基于自然场景统计(Natural Scene Statistics，NSS)的VQA模型、基于视觉感知的VQA模型和基于学习的VQA模型. 它们的具体介绍如下：

(1)NSS是感知领域的一门学科[7,65]，它涉及场景有关的统计规律[2]. 本质上，基于NSS的VQA模型是构建自然场景的统计规律(使用分布函数进行数学描述)，并量化失真对统计规律的影响. 常见的NSS包括亮度统计、颜色统计、空间相关性、高阶统计、时空统计等[65]. 该类模型不依赖于视频的失真类型，因而它们比针对特定失真设计的模型有更广的应用范围. 需要注意的是：自然场景[7]指的是光学相机所拍摄的自然界内容以及人造的室内/室外的场景，它用于区别人造图像(图形).

(2)基于视觉感知的VQA模型的思想是模拟HVS对视觉信号的感知过程，通过对HVS感知特性建模，构建与主观感知一致的VQA模型[5-6,66]. 常见的HVS特性包括：空间频率误差、亮度掩膜、纹理掩膜、空间频率误差敏感机制、短期记忆机制、时空池化等[67-68].

(3)基于学习的VQA模型主要指的是依赖深度神经网络(Deep Neural Network，DNN)的VQA模型[69-70]. 不同于前两类模型(依赖先验知识提取特征)，该类模型通过“学习”的方式自动获取视觉特征. 得益于优异的表征能力，基于DNN的VQA模型的性能优于基于浅层学习的VQA模型.

2.3 评价指标

常用的评价指标有三个，分别是皮尔森线性相关系数(Pearson Linear Correlation Coefficient，PLCC)、斯皮尔曼等级相关系数(Spearman Rank-Order Correlation Coefficient，SRCC)和均方误差(Root Mean Square Error，RMSE). 其中，PLCC用于计算预测准确性；SRCC用于计算预测单调性；RMSE用于计算预测一致性. 它们的计算方式如下：

$$PLCC=\frac{\sum_{i=1}^{N}(s_i-\bar{s})(p_i-\bar{p})}{\sqrt{\sum_{i=1}^{N}(s_i-\bar{s})^2\sum_{i=1}^{N}(p_i-\bar{p})^2}} \tag{4}$$

$$SRCC=1-\frac{6\sum_{i=1}^{N}d_i^2}{N(N^2-1)} \tag{5}$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{N}(s_i-p_i)^2} \tag{6}$$

式(4、5、6)中，s_i和p_i分别表示第i个视频的主观质量分数及客观质量分数，$\bar{s}$和$\bar{p}$分别表示主观质量分数平均值和客观质量分数平均值；N表示样本数量；d_i表示第i个视频主观质量分数排名与客观质量分数排名的差值.

如文献[71-72]所建议的，在计算PLCC和RMSE之前，需要对客观算法计算得到的分数进行非线性回归操作，包括四参数回归[71]和五参数回归[72]，它们的数学表达式如下：

$$g_1(p)=(\eta_1-\eta_2)\left[\frac{1}{1+\mathrm{e}^{\left(-\frac{p-\eta_3}{|\eta_4|}\right)}}\right]+\eta_2 \tag{7}$$

$$g_2(p)=\beta_1\left[\frac{1}{2}-\frac{1}{1+\mathrm{e}^{(\beta_2(p-\beta_3))}}\right]+\beta_4 p+\beta_5 \tag{8}$$

式(7、8)中，p表示原始的客观质量分数；η_1、η_2、η_3和

1 合成视频的一种，包括多视角拍摄的视频帧和合成帧
2 https://en.wikipedia.org/wiki/Scene_statistics

η_4为模型参数；β_1、β_2、β_3、β_4和β_5为模型参数. 常用的非线性回归操作是四参数回归，它可以保证预测结果的单调性.

计算PLCC、SRCC和RMSE这三个评价指标时依赖主观分数，而现有的包含主观分数的数据集中的视频图像数据相对较少，原因在于收集大量的主观分数是十分耗时耗力的. 众所周知，数据是驱动图像处理、计算机视觉等领域快速发展的动力，构建大规模数据集的高必要性和收集大规模主观标注的低可行性形成了矛盾. 2017年Ma等人[73]提出三个全新的评价指标，使用该指标时不需要主观分数. 这三个指标分别是基于序列排序一致性测试(L-test)、原始图/失真图可辨别性测试(D-test)和配对偏好一致性测试(P-test). L-test的目的是评估客观模型在对具有相同内容、相同失真类型但不同失真程度的信号进行评级时的鲁棒性，即性能好的客观模型能准确地区分相同内容、相同失真类型但不同失真程度的信号，失真程度越大质量越低；D-test是用于量化客观模型区分原始信号和失真信号的能力，即性能好的客观模型对于原始信号的预测分数应该高于失真信号的预测分数；P-test用于比较客观模型在质量可分的一对信号上的偏好预测，即性能好的客观模型能够对质量可分的一对信号给出一致性判断，对于质量好的信号的预测分数要高于质量差的信号的预测分数. 关于L-test、D-test和P-test的详细介绍可参照文献[43-44,73].

Ma等人[74]提出一种称为组最大差异化竞争(Group Maximum Differentiation Competition，gMAD)的评价指标. 该方法源于最大差异化竞争(Maximum Differentiation Competition，MAD)[75]，由Wang和Simoncelli于2008年提出，它的思想是“以合成的方式进行分析”(Analysis by Synthesis). 具体而言，给定一张图像$\boldsymbol{I}$和它的失真版本$\boldsymbol{I}_d$以及两个比较模型f_1和f_2，固定一个模型的结果然后优化另一个模型的最好和最差结果，通过比较优化得到的最好和最差结果的判别能力来判断模型的优劣. 如图2所示，两个模型分别是MSE和SSIM，图像A是对参考图像添加失真得到的；图像B和图像C与参考图像的MSE值相同，而SSIM值分别为最大(即最佳)和最小(即最差)；图像D和图像E与参考图像的SSIM值相同，而MSE值分别为最小(即最佳)和最大(即最差). 对比图像B和图像C，图像B的质量优于图像C的质量，SSIM的预测结果与主观感知更为一致，因而在该情况下SSIM优于MSE. 对比图像D和图像E，因为图像E包含多处局部的严重失真，质量更差. 在该情况下，MSE的预测结果与主观感知更为一致，因而表现更佳. MAD方法的不足包括：(1)对比的模型需要可导；(2)合成的图像不自然. gMAD方法很好地解决了上述问题，它将对比的模型扩展至任意的模型，不再需要模型可导，并且模型比较的样本是自然图像. gMAD的思想是在一个大的样本空间$\boldsymbol{D}$中寻找两两模型(如f_1和f_2)预测结果差异最大的样本，这类样本存在三种情况：(1)f_1和f_2均预测较好，这种情况当且仅当$\boldsymbol{D}$中的样本较少时存在；(2)f_1(或f_2)预测较好而另一个模型预测较差，这类样本可用于比较模型的优劣；(3)f_1和f_2均预测失败，这类样本虽然无法用于比较模型的优劣，但可以指出模型的不足和潜在的改进方向. 如图3所示，A和B为f_2预测结果一致而f_1预测结果差异很大的样本；C和D为f_1预测结果一致而f_2预测结果差异很大的样本. gMAD的思想也被扩展至图像分类[76]和图像语义分割[77-78]模型评估中，并被成功地用于IQA模型提升[79-80]和无偏测试数据的收集[81].

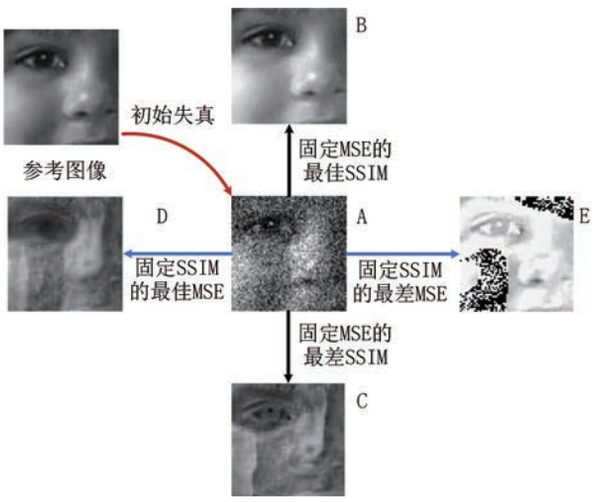


图2　MAD工作示例图[75]

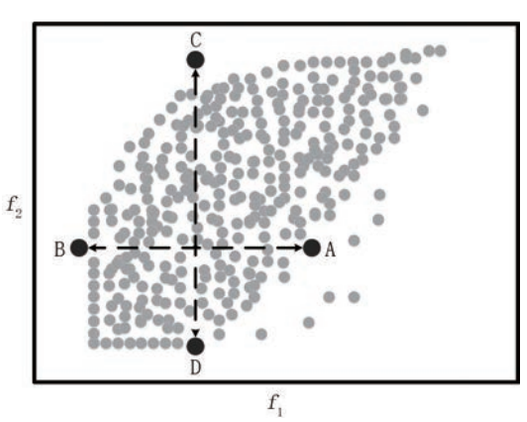


图3　gMAD工作示例图[74]

3 二维视频质量评价

3.1 数据集介绍

早期的二维视频质量主观数据集都是通过人工添加失真即合成失真(也称为模拟失真)生成的,模拟视频降质情况,常见的失真类型如压缩失真和传输失真等.随着多媒体技术的快速发展和广泛应用,真实失真逐渐成为该领域的研究热点.相对于合成失真,真实失真更加复杂多样,度量难度更大[44].失真视频帧示例如图4所示.主流的二维视频质量主观数据集如表2所示.

值得注意的是:(1)CSIQ数据集[84]主观实验采用SAMVIQ(Subjective Assessment Methodology for Video Quality)方法[1],该主观测试方法与其他主观测试方法最大的不同是受试者可以选择测试的顺序和更正他们的判断.(2)LSVQ数据集[54]中每个视频被裁剪成三种视频块,分别称为空域视频块(Spatial V-Patch, sv-patch)、时域视频块(Temporal V-Patch, tv-patch)和时空域视频块(Spatio-temporal V-Patch, stv-patch)[2],共117,225个视频块.其中,sv-patch指的是时域与源视频一致而空间分辨率为源视频的40%;tv-patch指的是空间分辨率与源视频一致而时域长度为源视频的40%;sv-patch指的是空间分辨率和时域长度均为源视频的40%.(3)KonVid-150k-A数据集[61]每个视频只被打分5次.

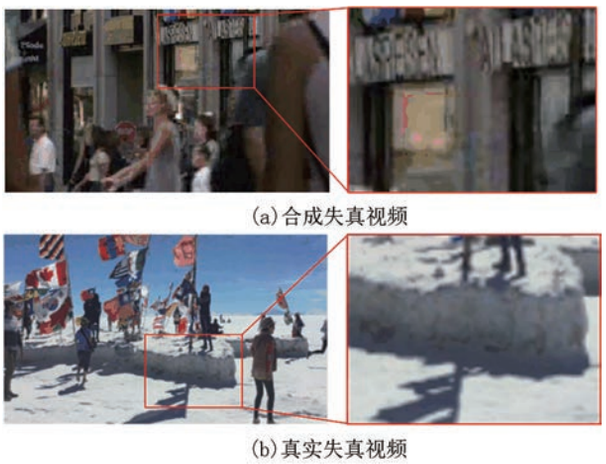


图4 合成失真视频和真实失真视频示例

从表2可知,2016年之后提出的数据集基本是针对真实失真的,数据量也从数百发展到十几万.目前已知的最大的数据集KonVid-150k-A[61]包含超过15万个视频,该数据集的缺点是每个视频只被打分5次,标签不够准确.该数据集的优点在于数据多,可用于模型的预训练[92].大部分数据集是通过众包的方式获取标签数据的,这是目前标注大量的

表2 二维视频质量主观数据集

名称	发表时间	源视频	失真视频	失真类型	视频时长(s)	分辨率	实验环境	主观方法	主观分数
EPEL-PoliMI[82]	2009	6	78	压缩失真	10	352×288	实验室	SSCQE	MOS
LIVE-VQA[83]	2010	10	150	压缩失真,传输失真	10	768×432	实验室	SSCQE	DMOS
CSIQ[84]	2014	12	216	压缩失真,传输失真	10	832×480	实验室	SAMVIQ	DMOS
MCL-V[85]	2015	12	96	压缩失真	6	1920×1080	实验室	PC	MOS
MCL-JCV[86]	2016	30	1,560	压缩失真	5	1920×1080	实验室	PC	-
CVD2014[87]	2016	-	234	真实失真	10~15	640×480/1280×720	实验室	SSCQE	MOS
LIVE-Qualcomm[88]	2018	-	208	真实失真	15	1920×1080	实验室	SSCQE	MOS
KoNViD-1k[89]	2017	-	1,200	真实失真	8	960×540	众包	SSCQE	MOS
LIVE-VQC[53]	2018	-	585	真实失真	15	<1920×1080	众包	SSCQE	MOS
YouTube UGC[90]	2019	-	1,500	真实失真	20	<3840×2160	众包	-	MOS
LSVQ[54]	2021	-	39,075	真实失真	5~12	1920×1080	众包	SSCQE	MOS
UGC-VIDEO[91]	2021	-	400	真实失真	10	1280×720	实验室	SSCQE	MOS
Youku-V1K[60]	2021	-	3,000	真实失真	10	1920×1080	众包	SSCQE	MOS
KonVid-150k-A[61]	2021	-	15,3841	真实失真	5	<960×540	众包	SSCQE	MOS
KonVid-150k-B[61]	2021		1,596	真实失真	5	<960×540	众包	SSCQE	MOS
PUGCQ[62]	2021	-	10,000	真实失真	5	<1920×1080	实验室	SSCQE	MOS

1 https://tech.ebu.ch/docs/techreview/trev_301-samviq.pdf.
2 此处缩写与原文保持一致

视频数据最有效的方式. 因为SSCQE的易操作性，它成为主观数据标注中最常用的方法[93].

3.2　模型介绍

如本文2.2小节描述的，本文根据VQA模型构建策略的不同将VQA模型主要分成三类(MSE和PSNR归为其他类，其中PSNR计算方式见公式(10))：基于NSS的模型、基于视觉感知的模型和基于学习的模型，详细分类如图5所示.

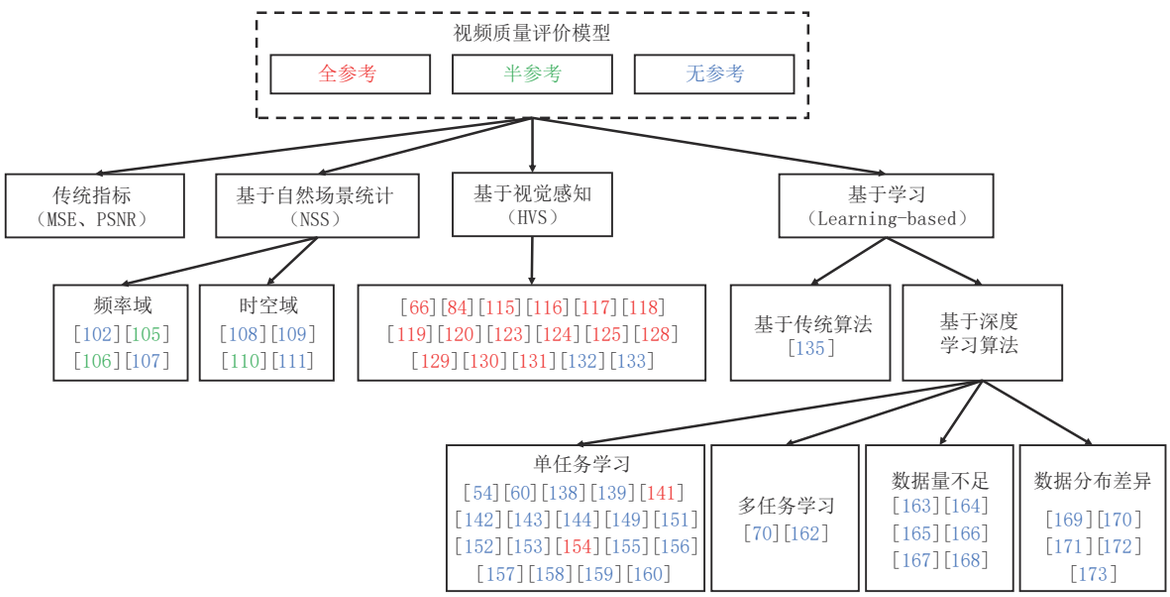


图5　二维视频质量评价模型分类

3.2.1　基于NSS的模型

根据NSS特征提取域不同，该类方法可以分成两类：基于频率域建模[94-97]的VQA模型和基于时空域建模[98-101]的VQA模型. 关于NSS的建模方式的介绍可参照文献[44].

(1)基于频率域建模的VQA模型. 该类模型的计算流程如图6所示，主要在帧差的基础上进一步获取视频的运动相关特征. Saad等人[102]在BLIINDS-II方法[97]的基础上提出V-BLIINDS，该方法包含三部分特征：基于帧差的二维离散余弦变化(2D Discrete Cosine Transform，2D-DCT)NSS特征、空间自然性分数和使用运动估计法[103]计算得到的运动特征. 基于半参考熵差模型(Reduced Reference Entropic Differences，RRED)[104]，文献[105]提出一种时空半参考熵差(Spatio-Temporal RRED，STRRED)模型，包含通过参考视频和失真视频单帧的傅里叶变换系数熵差计算得到的空域半参考熵差部分和通过参考视频和失真视频相邻帧差的傅里叶变换系数熵差

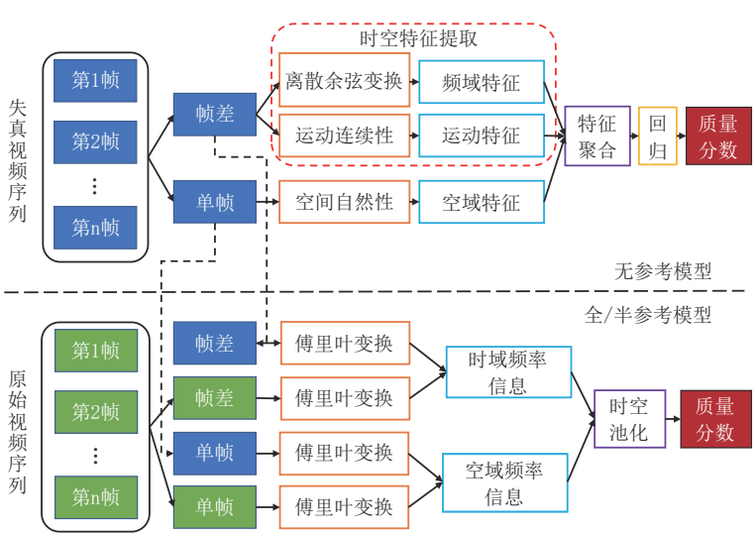


图6　基于频率域建模的VQA模型

计算得到的时域半参考熵差部分. Bampis等人[106]在RRED[104]和STRRED[105]的基础上提出SpEED-QA模型，建模方式与RRED和STRRED是一致的. Li等人[107]提出使用3D-DCT获取视频的NSS特征，包括谱特征、形状参数、能量波动和分布变化.

(2)基于时空域建模的VQA模型. 该类模型的计算流程如图7所示，该类模型主要依赖于自然性特征的提取，常使用的是非对称广义高斯分布(Asymmetric Generalized Gaussian Distribution，AGGD)模型，AGGD的数学表示如下：

$$f(x;\chi,\beta_l,\beta_r)=\begin{cases}\dfrac{\chi}{(\beta_l+\beta_r)\Gamma(\frac{1}{\chi})}\exp(-(\dfrac{-x}{\beta_l})^{\chi}), \forall x<0\\ \dfrac{\chi}{(\beta_l+\beta_r)\Gamma(\frac{1}{\chi})}\exp(-(\dfrac{x}{\beta_r})^{\chi}), \forall x\geqslant 0\end{cases} \tag{9}$$

式(9)中，χ控制分布的形状；β_l和β_r分别控制两边的扩散程度；Γ表示伽马函数. Mittal等人[108]提出使用空域自然性和运动自然性感知视频失真，该模型称为VIIDEO. 对于空域自然性，使用AGGD模型拟合帧差的局部归一化亮度图，将AGGD模型的参数表示空域自然性；对于运动自然性，首先拟合帧差局部归一化亮度图的低通图，计算相邻帧拟合分布参数的差表示运动自然性. 文献[109]提出使用3D亮度统计特征度量视频的时空信息变化；Yu等人[110]针对UGC视频压缩降质问题提出使用单帧亮度统计特征和相邻帧亮度差异统计特征感知UGC视频降质；Ebenezer等人[111]提出融合运动感知和空间信息NSS的VQA模型，称为ChipQA.

总体而言，NSS在VQA模型构建中取得了较大的成功. 该类模型需要依赖于视觉感知特征的分布先验，并使用分布的参数表示视觉感知特征. 而对于某些特定内容的视频可能不存在分布规律，因而研究者们提出了使用直方图统计等方法设计通用性更高的模型[44].

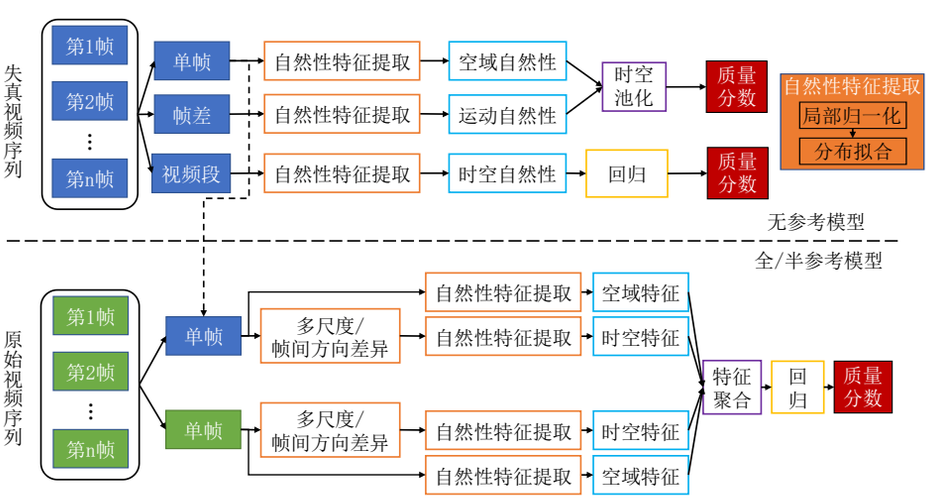


图7 基于时空域建模的VQA模型

3.2.2 基于视觉感知的模型

两个常用的VQA模型是MSE或者峰值信噪比(Peak Signal-to-Noise Ratio，PSNR)和SSIM[12]. PSNR和SSIM的计算方式如下：

$$f_{\mathrm{PSNR}}(F^r,F^d)=10\log_{10}[\frac{L^2}{\frac{1}{N}\sum_{n=1}^{N}(F^r-F^d)^2}] \tag{10}$$

$$f_{\mathrm{SSIM}}(F^r,F^d)=\frac{(2\mu_r\mu_d+c_1)(2\sigma_{rd}+c_2)}{(\mu_r^2+\mu_d^2+c_1)(\sigma_r^2+\sigma_d^2+c_2)} \tag{11}$$

式(10、11)中，F^r和F^d分别表示参考视频帧和失真视频帧；L表示视频帧像素点的最大动态范围，一般取值为255；N表示视频帧的像素点个数；μ和σ^2分别表示视频帧局部均值和方差；σ_{rd}表示参考视频帧和失真视频帧对应块的协方差；c_1和c_2为两个常数. PSNR以逐像素的方法计算视频的失真，并未考虑到HVS的感知特性. 而SSIM考虑了HVS的结构感知特性，能够获得与主观感知更为一致的预测结果. 类似于IQA模型使用空间池化策略[112-113]和多尺度策略[114]等以获得更好的性能，VQA模型经常需要结合时空权重策略来预测视频质量.

该类模型的计算流程如图8所示. 对于FR-和RR-VQA模型，它们往往先计算空间感知失真，然后通过加权策略获得全局失真；对于NR-VQA模型，它们先根据HVS特性提取感知特征，再通过回归模型进行训练. Wang和Bovik[66]较早地探索了HVS特性启发的VQA模型设计，充分考虑了空域和时域HVS特征，包括空间频率敏感性、亮度掩膜、纹理掩膜、时域频率敏感性和短期记忆效应. 该模型计算视频质量时将视频看成是单帧的集合，首先根据失真计算模型（如MSE）、HVS敏感特性和掩膜效应计算得到单帧的质量分数，然后根据时域频率敏感性对单帧分数处理；并根据短期记忆效应对帧序列分数进行平滑处理，最终将平滑后的帧序列分数的均值作为视频的分数. Wang等人[115]将SSIM算法引入到VQA中，提出一个三阶段（局部区域-单帧-视频）的VQA模型. Wolf和Pinson[116]提出一种VQA模型，称为VQM（Video Quality Model）. VQM包含7种独立的指标：空间信息丢失、水平或竖直至对角线方向的边缘偏移、对角线至水平或竖直方向的边缘偏移、颜色信息丢失、空间质量增强指标、运动边缘损伤和局部颜色损伤，将这7种独立的指标线性相加得到视频的质量分数. Li和Wang[117]提出基于运动感知的VQA模型，将运动感知分解为运动信息内容和感知不确定性，并将两者进行数学描述获得视频中任意时刻、任意空间位置的像素点的权重，通过结合IQA方法获得视频的质量分数. Ninassi等人[118]将视频降质定义成空域失真在时域的演变，将视频分割成短期视频段并计算时空失真，然后使用长期时空权重策略将时空失真图融合获得视频的全局分数.

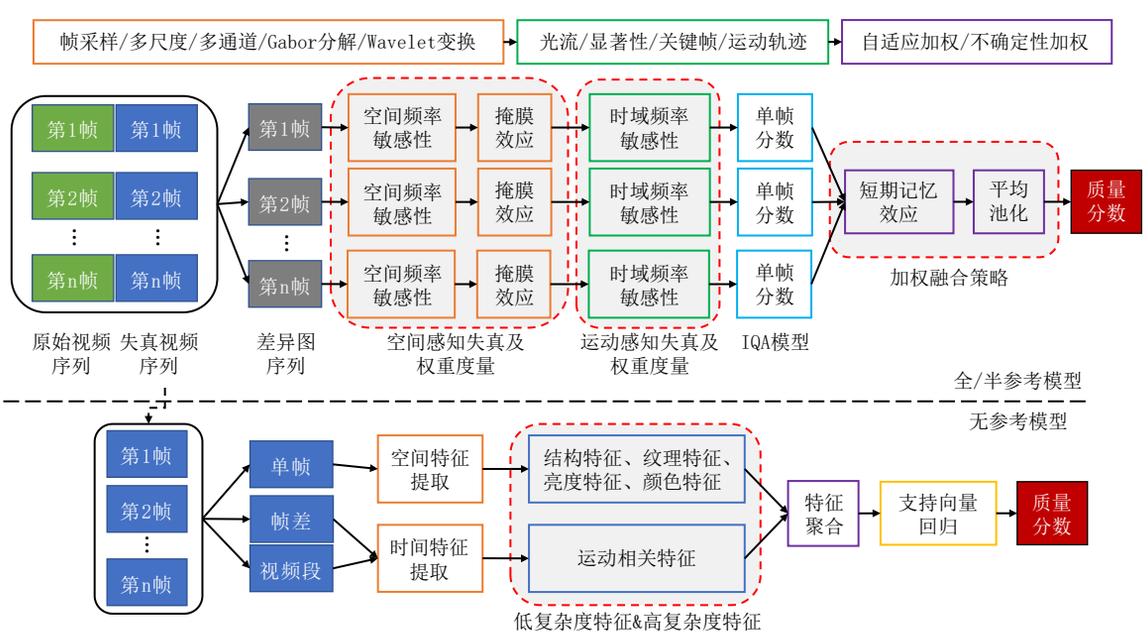


图8　基于视觉感知的VQA模型

考虑到视频运动表征质量对于视频质量感知是十分重要的，Seshadrinathan等人[119]提出融合空域、时域和时空域质量度量的VQA模型，称为MOVIE. 该模型主要依赖于Gabor分解，获取视频帧的多尺度表示. Lu等人[120]引入3D-Wavelet变换表征视频感知信息，模拟HVS的多通道感知结构；然后使用时空对比度敏感函数对3D-Wavelet系数加权，最后使用时域感知机制计算得到失真视频的质量分数. 美国公司Netflix提出视频多方法融合（Video Multimethod Assessment Fusion，VMAF）模型[1]，该模型首先提取像素级的空域特征和时域特征，其中空域特征包括视觉信息保真度[121]和细节及加性损伤[122]，时域特征为相邻帧亮度分量的差异；然后使用帧级的空间池化获得视频的特征表达. 然而，VMAF并未考虑时域掩膜效应. Bampis等人[123]在VMAF的基础上融合空域和时域方法，提出ST-VMAF和E-VMAF.

不同于上述方法，考虑到HVS特性、视频内容特点、播放设备属性和观看条件，加拿大公司SSIMWAVE[124]提出SSIMPLUS算法. 该算法在

1　https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652

多尺度下计算结构质量图，使用局部信息内容和失真对结构质量图进行空域加权得到所有尺度下的视频帧质量. 最后，根据HVS特性、视频内容特点、播放设备属性和观看条件计算视频分数. Zeng和Wang[125]将SSIM算法扩展成3D-SSIM算法并应用在VQA算法中. 类似于该研究[66]，考虑到视频中最严重的降质对整个视频的视觉质量有持续性的影响，作者[126]提出内容自适应的空域权重策略和时域权重策略，时空域区域质量越低对应的权重越大[127]. Vu和Chandler[84]提出联合空间失真度量和时空失真度量的VQA模型，该模型称为ViS$_3$. Manasa和Channappayya[128]提出使用局部光流统计来量化视频的时域失真和使用M-SSIM[114]计算视频的空域失真. 考虑到视觉记忆在视觉感知中的重要作用，作者[129]提出结合短期记忆和长期记忆模拟视觉记忆作用. 不同于研究[66,126]，该研究[129]使用显著图计算短期记忆和长期记忆. Korhonen[130]提出结合视频的时序特征、动态和空间失真相关统计特征的VQA模型. Wu等人[131]认为显著运动轨迹的捕捉是准确计算视频质量的关键，提出基于显著运动轨迹的VQA模型.

上述研究主要是针对合成失真视频，Tu等人[132]针对UGC视频质量问题在现有的UGC视频质量主观数据集上开展了系统性研究，发现通过简单地选择和聚合现有主流的VQA模型中的视觉感知特征就可以获得优异的性能. 作者还发现：(1)空域失真是UGC视频降质的主要因素；(2)使用运动相关特征度量移动设备拍摄的视频更加有效；(3)深度CNN特征可以有效地表征视频质量. Kancharla和Channappayya[133]发现UGC视频感知域下的运动轨迹直线度与视频质量是相关的，并提出使用预测运动轨迹误差度量视频的时域失真，最后结合空域失真度量得到视频的全局质量.

总体而言，这类VQA模型的研究重点在于将HVS特性转换成数学表达. 对于FR-和RR-VQA模型，它们的核心是度量单帧降质和计算时空域权重；对于NR-VQA模型，它们的核心是空域特征和运动特征的提取. 虽然现有VQA模型都注重时域特征/运动特征的提取，但研究[132]发现空域失真对视频质量的影响更大；该研究结论也得到了其他研究[25]的证实，将启发下一代VQA模型的设计.

3.2.3 基于学习的模型

不同于上述两类VQA模型，基于学习的VQA模型不需要人为设计特征，而是通过“学习”的方式自动获得视频质量相关的特征表达[69,70-134]，包括基于传统算法的模型和基于深度学习的模型. 它们的介绍具体如下：

(1)基于传统算法的模型. Xu等人[135]将CORNIA方法[136]中的图像质量特征学习方法扩展至视频质量特征学习，该方法称为V-CORNIA. 该方法首先学习单帧的空间特征表达，并使用IQA方法[137]计算单帧的客观分数，用于训练单帧视觉质量预测模型；然后，使用时域滞后权重策略融合单帧质量分数，得到视频质量. 该类方法的建模思想可参考文献[43-44].

(2)基于深度学习的模型. 如图9所示，根据模型的设计思路，将该类模型分成四小类：(1)单任务学习框架，指的是仅通过质量评价任务训练得到的模型；(2)多任务学习框架，指的是使用其他任务辅助质量预测主任务的模型；(3)解决数据量不足的模型，指的是通过预训练的方式获得更好的初始化权重的模型；(4)解决数据分布不一致问题的模型，指的是使用域迁移方式学习更加一般性的视频质量特

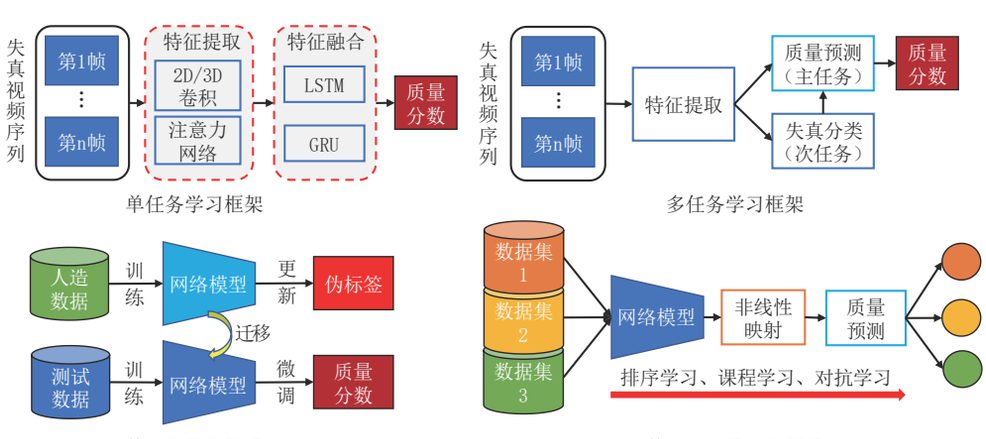


图9 基于深度学习的VQA模型

征的模型.

第一类是单任务学习框架. 类似于视频分类等模型的设计，该类模型的研究重点在于时空特征融合. Wang等人[138]较早地将3D-CNN引入VQA模型中. 随后，Wang等人[139]提出基于时空融合的VQA模型，其输入为一个IQA模型[140]提取的特征和相邻帧的SSIM图均值以及方差向量的均值和方差. 考虑到时域运动效应和时域记忆效应，Kim等人[141]提出基于两阶段训练的VQA模型，两阶段分别用于模拟时域运动效应和聚合视频帧分数. Zhang等人[142]提出基于弱监督学习和重采样策略的VQA模型，核心是使用M-SSIM[114]生成视频块的弱监督学习标签数据预训练模型. Li等人[143]提出时域记忆机制启发的VQA模型，称为VSFA. 该模型使用ResNet50[92]获得视频帧的语义特征表达，使用门控循环单元(Gated Recurrent Unit，GRU)获取帧间的时序信息. Chen等人[144]考虑运动感知机制提出融合多种运动频率的VQA模型，称为RIRNet. 与VSFA不同的是：RIRNet考虑不同频率采样的时序信息，使用空间金字塔池化(Spatial Pyramid Pooling，SPP)层[145]保证每张特征图输出固定长度的特征向量；使用密集监督信号辅助网络的训练. Ying等人[54]提出一个双通道时空特征融合的VQA模型，称为PVQ. 该模型的双通道包括一个预训练好的IQA模型PaQ-2-PiQ[146]和一个预训练好的动作识别网络3D ResNet-18[147]，并使用InceptionTime模型[148]作为质量回归模型. Yi等人[149]提出一个基于注意力机制的VQA模型，该模型使用VGG[150]作为视频帧特征提取网络，并在VGG的第8、9和10层插入注意力层，然后将特征提取网络的输出作为GRU学习帧间时域信息，使用时域记忆机制启发的权重策略融合单帧分数得到视频分数. 考虑到HVS更加关注失真区域，Chen等人[151]设计了一个时空聚合网络，用于自适应学习各时空块的权重，以便融合各时空块质量得到视频的质量. Zhu等人[152]提出一个端到端的VQA模型，包含空域特征提取模块、局部运动特征提取模块和时域质量聚合模块. 为了解决在处理高分辨视频时需要复杂计算和高存储要求的问题，Wu等人[153]提出网格小块采样策略，其核心思想是用这些在空域随机分割和时域对齐的小块表示整个视频.

考虑到互联网视频的特点，Xu等人[60]提出一个时空失真感知模型，包括帧级子模型和视频级子模型. Liu等人[154]提出基于连续依赖建模的VQA模型，该模型包括单帧降质计算、连续依赖降质建模和质量预测模块. Wu等人[155]提出一个融合高层级语义特征和低层级特征的VQA模型，该模型包括两个分支分别用于处理高级语义特征和低级特征. Shen等人[156]提出一种基于层级时空特征表达的VQA模型，该模型包括多尺度特征提取模块、层级时空特征融合模块和质量回归模块. 该模型的特点在于质量回归模块将不同尺度的特征映射为主观分数以及将不同阶段获得的融合特征映射为主观分数，不同的是融合特征经过注意力模块自动学习不同尺度特征的重要性. 受“分而治之”思想启发，Li和Yang[157]提出分层级自注意力网络，其核心是渐进式地获取单帧、视频切片和整个视频的质量表达. 类似于该工作[157]，You和Lin[158]使用多头注意力机制融合每个视频切片的特征进而得到视频的全局特征表达，最后用于视频质量的预测. Xing等人[159]提出一种时空注意力模型用于VQA任务，该模型的编码模块为多个串联的时域注意力子模块和空域注意力子模块. 为了捕捉视频序列长期时空依赖，该模型将每个块的时空位置信息编码后作为输入. 考虑到HVS中刺激驱动的自底向上注意力机制和感知驱动的自顶向下注意力机制是共同作用的，Guan等人[160]构建一个视觉注意力模块和一个记忆注意力模块，分别提取帧级和视频级质量特征，最后将视频级特征用于质量预测.

第二类是多任务学习框架. Liu等人[70]提出多任务学习的VQA模型，称为V-MEON. 该模型是在一个IQA模型MEON[161]的基础扩展的，包括两个分支：一个分支用于预测视频的编码类型；另一个分支用于预测视频的质量. Wang等人[162]提出使用内容信息、失真信息和压缩程度感知视频失真，提出一个三通道的VQA模型，三个通道对应地输入为单视频帧、单视频帧和多视频帧. 该模型将三个通道输出的特征聚合，用于预测视频的质量.

第三类是解决数据量不足问题的模型. Liu等人[163-164]提出基于时空表达学习的VQA模型，包括特征编码模块和分层特征回归模块. 为了解决质量数据少的问题，作者收集了超过32万个视频，并使用4个客观算法生成视频的伪标签，用于预训练模型. Chen等人[165]首次将对比自监督学习(Contrastive Self-Supervised，CS)引入VQA研究，提出基于CS预训练(Pre-Training)(CSPT)的VQA框架. 为了更好地表征视频质量信息，作者提出失真相关对比学习和内容相关对比学习. 同时，将失真类型判别

任务用于增强视频质量特征的学习能力. Li等人[166]提出使用图像质量主观数据集和动作识别数据集分别训练空域特征提取器和时域特征提取器,然后将空域和时域特征联合输入GRU模块用于预测视频质量. Mitra和Soundararajan[167]将自监督多视图对比学习引入VQA任务中,用于学习视频的时空质量特征表达;然后使用多元高斯(Multivariate Gaussian, MVG)模型计算失真视频与质量完好视频特征表达之间的距离而获得视频质量. 不同于该工作[165],Jiang等人[168]将失真分类、帧率预测和比特率差异预测任务引入到CS框架中以更好地学习视频质量特征.

第四类是解决数据分布不一致性问题的模型. 考虑到不同的数据集的主观分数是有差异的,Li等人[169]提出跨数据集的VQA模型训练策略. 该研究工作使用三个损失函数,包括单调性损失函数、线性损失函数和误差损失函数. 其中,单调性损失函数类似于成对损失函数,保证模型预测排序的一致性;线性损失函数指的是预测分数和主观分数的PLCC;误差损失函数指的是L_1范式. Chen等人[170]提出无监督域迁移的VQA模型,包括域自适应质量预测模块、不确定性排序模块和自监督子域调整模块. 其中,域自适应质量预测模块包括基准模型、分数预测模型和域分类器,将源域和目标域的特征分布对齐;不确定性排序模块将目标域分成确定性和不确定性子域;自监督子域调整模块用于消除确定性子域和不确定性子域的差异. 为了解决数据分布不同导致模型测试性能很差的问题,Chen等人[171]提出基于可泛化时空特征表达的VQA模型. 该模型首先使用预训练好的VGG提取单帧的多尺度特征,并使用注意力模块进一步地处理多尺度特征;然后通过对抗学习将帧级别的特征约束为高斯分布;最后使用金字塔池化聚合视频时域质量. 针对VQA模型可扩展能力弱和有效性较低的问题,Xian等人[172]提出融合迁移学习和生成伪参考视频策略的解决方案. 该方案将输入视频及其伪参考视频的残差视频叠加后作为VQA模型的一部分输入,输入视频经过生成模型处理后的层级特征作为另一部分输入;该VQA模型使用3D卷积和门控循环单元融合不同层级的特征,然后作为质量预测模块的输入. 考虑到不同源域的专业知识有助于提升模型的可扩展能力,Chen等人[173]提出基于动态专业知识集成的VQA模型. 作者首先在每个源域训练一个专家模型;然后在对比学习的框架下训练集成模型;最后将目标域与各个源域的相关度作为权重用于融合各个专家模型的输出,形成聚合的特征并用于质量预测[173].

总体而言,基于学习的VQA模型主要是通过端到端的学习获取视频与质量之间的映射关系. 其中,第一类模型着重于提取连续帧之间有效的感知质量特征,并通过非线性映射得到视频质量;较第一类模型相比,第二类模型通过引入额外的监督信号提升模型的表征能力;第三类和第四类模型分别针对数据量不足和数据分布不一致性问题,着重解决模型的扩展性问题,这也是当前VQA研究的重点. 得益于优异的数据表征能力,基于学习的VQA模型的性能明显优于其他类型模型.

3.3 小结

从视频质量主观数据集的角度来看,失真类型从单一失真到多种失真转变,均是面向应用场景开展的研究工作. 同时,数据集的规模也从百级发展到万级. 从客观模型的角度来看,得益于深度学习强大的表征能力,基于深度学习的客观模型逐渐成为研究的重点. 除了研究更具表征能力的VQA模型,研究者也开始着力于解决VQA模型跨域问题和提升模型的可扩展能力.

4 其他内容视频质量评价

4.1 主观数据集介绍

本章介绍的其他内容视频包括:三维视频、屏幕内容视频、合成视频和全景视频. 三维视频包括左右视频,与二维视频相比三维视频多一维深度信息[174],观看三维视频时存在双目视觉现象;屏幕内容视频指的是显示电子设备桌面内容的视频,内容由图像/图形部分和文本部分组成. 屏幕内容视频视觉特征分布与二维视频视觉特征分布差异较大,并且不同部分的内容对视觉感知的影响存在差异[175-176];合成视频主要指的是通过基于深度图的绘制(Depth Image based Rendering, DIBR)技术生成的虚拟视点视频,与二维视频的区别包括:这类视频可以提供任意的观看视角,而二维视频只能提供固定的视角;合成视频失真呈现非局部均匀性[177-178];全景视频(也称为360°视频)包含任意视角,由多个摄像头拍摄后渲染而成,是虚拟现实沉浸式多媒体主要的数据形式,在某一个时刻用户只能观看到一个视口. 它们与二维视频的区别如图10所示,不同类型视频帧如图11所示,当前主流的不同内容视频

的质量主观数据集如表3所示．相对于二维视频质量主观数据集，其他内容视频质量主观数据集数据较少，最大的数据集[25]包含1，944个视频．

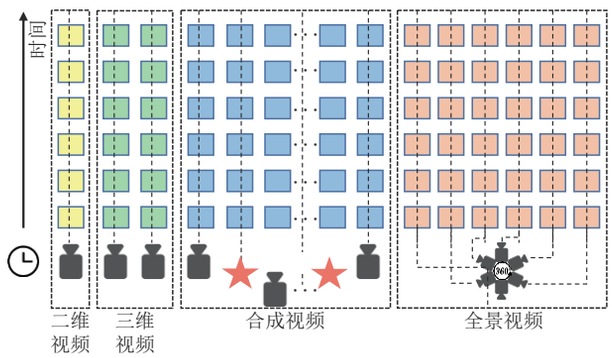


图10　不同类型视频对比（黑色相机表示真实相机，红色五角星相机表示虚拟相机）[25]

图11　不同类型视频示例

需要补充说明的是：与通过平面显示器可以直接观看的其他类型视频不同，全景视频通常需要借助特定的投影转化才能正常播放观看．目前，针对全景视频的编码压缩方式依托于普通二维视频的框架，然而，由于两者空间维度的不一致，3D球面视频无法直接被2D编码框架处理，全景视频需要先通过一定规则的投影映射至2D平面才能进行编码、压缩与传输，在播放观看时进行解码重投影返回到3D球面空间．球面投影多为非线性映射，不同的投影格式会给视频带来不同程度的形变，影响最终以球面状态呈现的视觉质量，因此投影格式的优化和选择同样也成为了质量评价和全景视频压缩编码中的重要内容．现在主流的投影方式有等距柱状投影（Equi-Rectangular Projection，ERP）、立方体投影（Cubemap Projection，CMP）、截断金字塔投影（Truncated Square Pyramid Projection，TSP）和条带投影（Segmented Sphere Projection，SSP）等．以下简要介绍主流的全景视频投影格式：

（1）ERP．该投影格式是目前应用最广泛的一种投影格式．ERP算法实现复杂性较低，空间连续性较好，几乎各大全景视频播放平台和头戴式显示设备（Head Mounted Display，HMD）都支持ERP转换．ERP的转换过程可以理解为将立体地球仪展开为平面地图的过程．ERP要求每条纬线上拥有相同的采样点数，保证2D展开平面的条件下，对每行像素进行了不同比例的拉伸．从赤道到两极，拉伸的比例越大，需要引入的冗余像素也越多，这样极大地影响了编码效率．从图12中可以看到高纬度区域扭曲的现象十分明显．

（2）CMP．该投影格式是一种常见的立方体贴图投影，它主要通过透视的方式完成球面到立方体面的映射．CMP的实现过程可以理解为将球面上内容投影到外切立方体后，将这个立方体以一定的规则展开成平面的过程．与ERP相比，CMP的每个面上的像素相对均匀，在一定程度上避免了两极过度拉伸引入冗余像素而造成编码低效的情况，但是不同的展开规则都不可避免地破坏了视频在空域上的连续性，容易在编码过程中出现匹配错位的问题．图13展示了CMP的两种排列形式，(a)中显示的非紧凑式排列，存在大量的冗余像素（灰色部分），(b)中显示的紧凑式排列，是将索引面重新排列，将编号为3、1、2的部分拼接在4、0、5的下面，这种CMP排列方式又称为重建立方体投影（Reshaped Cubemap Projection Format，RCMP）．

（3）TSP．该投影格式又称为四棱台投影格式，是一种新型的投影格式．与传统投影格式中每个面都相同的特点不同，TSP有选择性地对棱台的6个面进行了不同下采样操作．如图14所示，以棱台底面为用户的主要观看视区，进行原分辨率采样和投影；以棱台的4个侧面为用户的次要观看视区，进行低分辨率采样和投影；以棱台顶面为用户不可观看视区，进行最低分辨率采样和投影．该种可变式的投影策略可在保证用户观看视频质量的同时忽略掉无关像素，从而减轻带宽压力．在实际应用中，使用TSP投影格式需要提前生成多种四棱台以供用户选择不同关注方向的内容．

（4）SSP．该投影格式是一种分割类的投影，实现

表3 其他内容视频质量主观数据集

名称	发表时间	源视频	失真视频	失真类型	视频时长(s)	分辨率	实验环境	主观方法	主观分数
[1]SCVD[179]	2020	16	800	高斯噪声,高斯模糊等	10	1920×1080	实验室	DSIS	MOS
[1]CSCVD[180]	2020	11	165	压缩失真	10	1280×720	实验室	SSCQE	MOS
[2]LIVE 3D[181]	2012	6	54	压缩失真	13, 15	720×480	实验室	SSCQE	DMOS
[2]NAMA3DS1-COSPAD1[182]	2012	10	100	H.264压缩,边缘增强等	13～16	1920×1080	实验室	ACR	MOS
[2]StSD 3D[183]	2013	14	116	压缩失真	8	1920×1080	实验室	DSCQS	DMOS
[2]Waterloo-IVC Phase I[184]	2017	4	176	压缩失真等	6, 10	1024×768, 1920×1080	实验室	ACR	MOS
[2]Waterloo-IVC Phase II[184]	2017	6	528	混合编码,量化编码等	6, 10	1024×768, 1920×1080	实验室	ACR	MOS
[2]LFOVIAS-3DPh2[185]	2019	12	288	压缩失真等	5, 7, 9	1920×1080	实验室	ACR	DMOS
[3]Bosc11[186]	2011	3	84	DIBR算法	6	1024×768	实验室	ACR	MOS
[3]Bosc13[187]	2013	6	276	DIBR算法	6	1920×1080, 1024×768	实验室	ACR	MOS, DMOS
[3]SIAT[188]	2015	10	140	压缩失真	6, 8	1024×768, 1920×1088	实验室	ACR	DMOS
[3]IPI-FVV[189]	2019	3	120	DIBR算法	5, 10	1920×1080	实验室	ACR	MOS, DMOS
[3]Youku-FVV[25]	2022	18	1,944	压缩失真	7, 9, 11	1920×1080	实验室	AccAnn	MOS
[4]Singla17[190]	2017	6	60	压缩失真	10	1920×1080～3840×2160	实验室	SSCQE	MOS
[4]Curcio17[191]	2017	3	24	压缩失真	21	3840×1920	实验室	SSCQE	DMOS
[4]Tran17[192]	2017	3	60	压缩失真	30	1440×720～3840×1920	实验室	SSCQE	MOS
[4]IVQAD[193]	2017	10	150	MPEG压缩,下采样	15	1024×512～4096×2048	实验室	SSCQE	MOS
[4]Zhang17[194]	2017	16	384	压缩失真	10	4096×2048	实验室	SSCQE	MOS
[4]Zhang18[195]	2018	10	50	H.265压缩,下采样	10	3600×1800	实验室	SSCQE	DMOS
[4]Lope18[196]	2018	6	240	H.265压缩,下采样	10	960×480～7680×3840	实验室	SSCQE	MOS
[4]VQA-ODV[197]	2018	60	540	压缩失真,投影失真	10～23	3840×2160～7680×3840	实验室	SSCQE	MOS
[4]VOD-VQA[198]	2021	18	774	H.264压缩,下采样	15	320×240～1280×960	实验室	SSCQE	MOS
[4]Zou21[199]	2021	10	354	传输失真	20	3840×2160	实验室	SSCQE	MOS
[4]JVQD[200]	2021	15	45	HEVC压缩, 抖动	6, 7, 8	4096×2048	实验室	SSCQE	MOS

注:1 屏幕内容视频;2 三维视频;3 合成视频;4 全景视频.

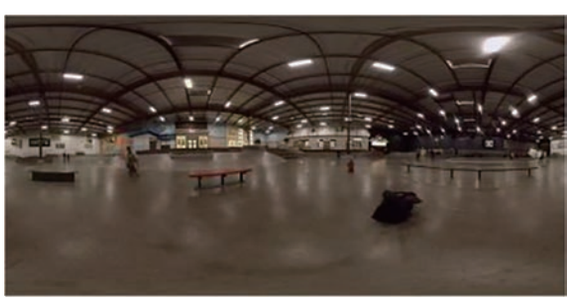
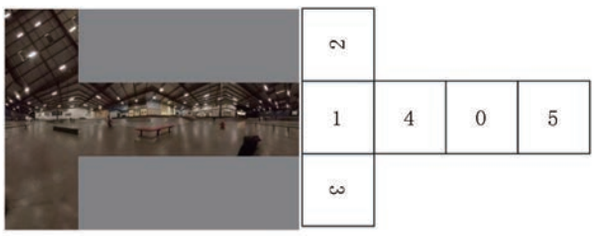


图12 ERP投影

思路是以北纬45°和南纬45°为边界将球面划分成为北极区域、赤道区域和南极区域3个部分.两极区域的像素投影在圆平面上,赤道区域的像素投影在条状的矩形平面上,然后将三者拼接起来.图15展示SSP投影格式,该种投影格式在赤道区域采用了均匀性高的投影方式,而在两极区域采用了均匀性较低但效率高的圆面投影,兼顾了观看质量和投影的复杂程度.



(a)排列方式1

(b)排列方式2

图13　CMP投影

图14　TSP投影

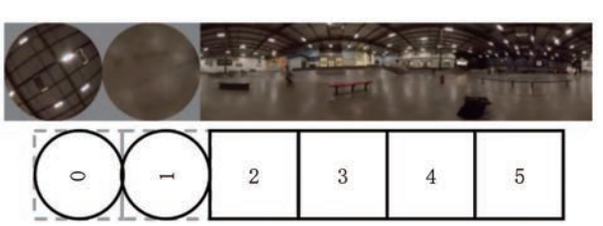


图15　SSP投影

4.2　模型介绍

4.2.1　屏幕内容视频

针对该类视频设计的VQA模型主要考虑屏幕内容特点. Cheng等人[179]较早地提出使用3D-Gabor滤波获取屏幕内容视频的时空信息，并用于计算屏幕内容视频的空域和时域相似性. Li等人[180]提出一种多尺度相对标准差相似性方法，该方法在帧差图的基础上计算相对标准差多尺度相似性；最后使用平均池化计算得到测试视频的分数. Li等人[201]认为屏幕内容感知应该基于多维度特征，并提出融合帧内特征和帧间特征表示屏幕内容视频质量. 其中，帧内特征包括梯度统计特征、标准差比例特征、压缩特征和频率域特征；帧间特征通过计算帧差的均值得到. Zeng等人[202]提出使用3D高斯拉普拉斯算子和3D NSS分别度量失真屏幕内容视频及其参考视频的相似性，并根据局部视频活动计算得到两部分相似性的融合权重，进而获得整体的质量.

4.2.2　三维视频

最直接的方法是使用2D-VQA算法分别计算左右视频的质量分数[203]，然后使用平均权重得到三维视频的质量. 研究[204-205]表明，因为双目视觉的存在，直接平均左右视频的质量分数得到三维视频整体的分数在非对称失真的情况下无法获得较好的性能. 为了有效地表征三维视频的视觉质量，研究者们陆续地提出一系列模型. 类似于二维视频质量评价模型的介绍，我们将三维视频质量评价模型也分成三种，具体介绍如下：

(1)基于NSS的模型

Appina等人[206]使用二元(Bivariate) GGD (BGGD)模型建模运动信息和深度信息的联合分布，通过计算多尺度、多方向的联合分布协方差矩阵的特征值，得到帧级别一致性分数. 另外，作者使用2D-IQA算法计算左右视频单帧质量分数；然后通过均值策略计算得到三维视频的空间质量分数；最后，融合一致性分数和空间质量分数得到三维视频的质量分数[206]. Jiang等人[207]首先对三维视频做向量分解处理以获得三维视频的运动特征图，在此基础上使用局部归一化操作并使用GGD和AGGD模型拟合局部归一化操作得到的特征图，使用GGD和AGGD模型的参数作为感知三维视频质量变化的一部分特征；此外，计算局部空间熵、谱熵以及谱熵直方图的均值和偏态，作为感知三维视频质量变化的另一部分特征. Chen等人[208]引入AGGD模型用于建模左右视图叠加图的分布，并将分布的参数感知三维视频的质量. 考虑到视觉掩膜效应和运动感知机制，Yang等人[209]提出基于关键帧检测的三维视频质量评价模型. 该模型首先通过帧差计算运动幅值，根据运动幅值确定关键帧；然后计算左右视频帧的叠加图和差异图，使用局部二值模式(Local Binary Pattern，LBP)提取叠加图和差异图不同方向的统计特征，之后使用PCA降维形成三维视频最后的特征表达. 在该文献[210]中，作者提出使用二元GGD模型获取三维视频运动信息和视差信息的联合统计分布，并使用MVG模型计算测试三维视频和高质量三维视频的MVG模型的差作为度量其质量的一部分.

(2)基于视觉感知的模型

考虑HVS的选择性处理机制和恰可识别机制，Qi等人[211]提出使用恰可识别计算失真三维视频和参考三维视频的相似性；然后模拟选择性处理机制计算帧内显著性、帧间显著性和双目显著性得到三

维视频的显著性图；最后融合相似性和显著性图得到三维视频的质量分数. Yang等人[212]将三维视频质量感知分成三部分，包括空域质量感知、时空域质量感知和时域质量感知. 其中，空域质量感知部分包括使用LBP提取的叠加图和差异图的分布特征以及多尺度多方向Log-Gabor变换特征图的幅值、方差和熵特征；时空域质量感知首先计算左右视频交叉帧差图，然后计算帧差图的幅值、方差和熵特征；时域质量感知使用光流法计算得到的运动方向和速度特征表示. 考虑到时序双目竞争机制，Fang等人[213]提出一个两阶段的方法. 该方法分别使用一个2D-IQA方法计算单帧空域失真和运动向量幅值差计算时域失真，然后使用结构强度和运动能量融合空域和时域的质量. 最后，提出一个时序双目竞争机制启发的权重策略融合左右视频分数得到三维视频质量分数. 在该文献[214]中，Fang等人提出基于感知视觉信息的权重策略，并系统性地研究了不同权重策略结合2D-IQA与VQA算法在三维视频质量评价的应用.

(3)基于学习的模型

Zhou等人[215]较早地提出一个端到端学习的双通道深度神经网络(End-to-end Dual Stream Deep Neural Network，EDN)框架，该框架包括权值共享的特征提取网络、全连接网络和质量回归模型，输入是左右视图的图像块. Feng等人[216]提出基于注意力机制和3D-CNN的模型，该模型使用双通道分别提取左右视频的多尺度信息，并使用包含不同膨胀率的多尺度单元和注意力模块逐渐地融合左右视频信息，最后使用3D卷积网络进一步获取三维视频的时序信号，通过全连接网络输出分数. 类似地，Yang等人[217]将3D-CNN引入三维视频质量评价模型中，该模型的输入为左右视频的差异视频.

4.2.3 合成视频

不同于二维视频和三维视频的失真，合成视频的失真主要源于DIBR过程引入的非均匀失真和时域不连续失真. 因此，针对合成视频设计的VQA方法的设计思路主要在于捕捉这两种类型的失真. 具体介绍如下：

Liu等人[188]提出了一种基于时空活跃度计算和时序闪烁度量的合成视频质量评价方法，该方法使用时空活跃度合成视频的模糊和块效应，使用时序梯度向量度量时序闪烁造成的失真. Sun等人[218]提出计算块内容亮度失真和对比度失真来度量帧失真，同时计算块边缘失真用来度量视频帧变化. Kim等人[219]提出首先检测合成视频中的闪烁区域，然后计算相邻帧闪烁区域的结构相似度作为帧级失真；最后将单帧闪烁区域像素点个数与所有帧闪烁区域像素点个数的比值作为单帧的权重，用于融合帧级失真得到视频质量分数. Huang等人[220]使用几何失真度量和时空不连续度量共同计算合成视频质量. Zhang等人[221]提出使用稀疏表示度量合成视频的时域闪烁失真，并联合IQA算法计算得到的空域失真预测合成视频质量. Stankovic等人[222]考虑DIBR过程引入的边缘失真，引入形态学多尺度计算用于预测合成视频质量. Ling等人[223]考虑到FVV中非均匀时空失真，提出首先定位显著运动路径，然后通过计算时域的结构失真得到FVV QoE分数. Zhou等人[224]认为闪烁即时序不连续是合成视频质量主要的影响因素，提出使用块梯度变化检测闪烁区域，然后使用奇异值分解计算闪烁区域失真，最后融合所有帧的失真分数得到合成视频的质量分数. 考虑到DIBR过程引入的几何失真会增加虚拟视点合成帧的高频信息，Wang等人[225]提出使用高频信息能量度量空域失真和连续帧的运动差异来量化时域不连续性，以度量时域失真. 最后，联合空域失真和时域失真计算合成视频质量.

除了上述基于传统算法的模型，研究人员也将深度学习技术引入到合成视频质量评价模型中. 考虑到FVV QoE主观数据集中数据量少的问题，Ling等人[226]提出使用生成对抗网络(Generative Adversarial Network，GAN)生成数据. Yan等人[25]提出基于稀疏采样的FVV QoE评价模型. 该模型的主干网络是一个性能优异的VQA模型(VSFA[143])，包括特征提取模块和时空特征融合模块. VSFA的原始输入为所有视频帧，而Yan等人[25]通过大量实验发现少量帧依然可以准确地预测FVV QoE.

4.2.4 全景视频

目前，围绕全景视频开展研究的客观质量评价方法主要分为三类：基于PSNR/SSIM改进的方法、基于视觉感知的方法和基于学习的方法. 具体介绍如下：

(1)基于PSNR/SSIM改进的方法

在针对全景视频设计的客观质量评价方法中，早期方法的设计思路大多来源于PSNR和SSIM，包括：基于全景视频特点改进的S-PSNR(Spherical PSNR)[227]、WS-PSNR(Weighted to Spherically uniform PSNR)[228]、V-PSNR(Viewport PSNR)[229]、

S-SSIM(Spherical SSIM)[230]等. S-PSNR通过在球面上均匀采样获得固定的采样点数来计算PSNR，从赤道到两极区域，球面上每行的采样点数随着纬度的增加而减少. 在360Lib[1]中，S-PSNR提供了球面上的经纬度坐标，通过空间投影关系可以获取参考和失真2D平面图像上的像素坐标，然后对其进行PSNR计算. 通常，在获取2D平面上的采样点坐标时会面临取到非整数坐标位置的情况，S-PSNR-NN(S-PSNR at Nearest Neighbor)则会对其直接进行取整操作，这也是目前大部分计算S-PSNR会选择采用的计算方式. 虽然S-PSNR改进了PSNR的非均匀采样问题，但是S-PSNR的固定采样点数还不足支持分辨率高于2K的典型全景图像的像素数量；因此在高分辨率全景图像上，S-PSNR难以获得较为准确的质量估计. WS-PSNR考虑了从球面投影转换到2D平面上的每个像素点对整体质量的影响与其在球面上对应的面积有直接关系，在基于2D平面计算PSNR的情况加入了与面积相关的权重因子. WS-PSNR可以直接在2D平面上操作，计算复杂度较低；但是因为不同的投影格式有不同的权重因子，所以WS-PSNR并不能直接用于跨投影格式的质量比较. V-PSNR通过生成与头部运动数据相对应的视口来计算视口之间的PSNR. 由于实际情况的头部运动数据是事先未知的，V-PSNR使用不同观看方向的可能性来近似模拟头部运动数据，比如，人们可能更倾向于观看赤道周围的区域而忽略两极区域内容. 相应的视口是一系列与球面相切的平面，由投影转换即可得到. V-PSNR提出了基于视口的计算方式，更符合人类实际观看情况，但是真实的头部运动数据可能更为复杂，仅仅借助观看方向的概率分布可能丢失时域上的信息分布. 与S-PSNR相似，S-SSIM[230]将SSIM中的对比度、亮度及结构相似性的计算放在了球面上操作，其采样过程也变成了在球面上对像素选取滑块窗口. 同时，考虑像素在球面投影转换到2D平面上的面积变化，S-SSIM以面积缩放比例作为权重来消除投影形变带给质量评价的影响.

(2)基于视觉感知的方法

该类方法主要考虑用户观看全景视频过程中的局部视口感知特性和视觉敏感特性，结合局部视口和视觉敏感特性建模度量全景视频质量. Xu等人[231]基于对主观实验数据的分析，提出了一种基于无内容感知的质量评估(NCP-PSNR)方法. 他们基于数据库获得了头部运动(Head Movement，HM)的经纬度分布，以生成HM的权重分布，将其用于PSNR的权重计算[231]. 此类方法运用HM的统计分布作为权重分配的感知依据，是一种常见的权重分配手段，但是它无法很好地模拟特定全景视频的注意力权重. 因此，Xu等人[231]提出了一种基于内容信息的质量评估(CP-PSNR)方法，运用随机森林模型预测每个视频的注意力分布，以此作为计算PSNR的权重分布. 针对全景视频中目标运动引起的晃动问题，Mahmoudpour等人[200]提出基于视觉掩膜估计和目标跳变率函数的全景视频质量评价方法. 其中，视觉掩膜估计是考虑运动目标不同的运动速度对失真可察觉度的影响；目标跳变率函数用于量化人眼追踪过程中不同速度的目标在错误空间位置跳变的频率. Jiang等人[232]提出基于分块的全景视频质量评价方法，该方法使用VMAF计算每一个投影至二维平面的块的质量分数；并根据HM数据和眼部运动(Eye Movement，EM)数据生成权重图，最后使用权重图加权块的分数得到全景视频的分数. 进一步地，考虑到基于块的全景视频传输系统中的网络变化和视口预测偏差，作者提出融合编码参数、视频内容和视口位置的块质量损失模型[232]. Gao等人[233]提出一种基于时空失真建模的全景视频质量评价方法，其核心思想是模拟HVS的注意力机制并设计相应的全景视频质量评价模型. 该方法包括四个步骤：(1)生成时空片段；(2)计算单片段的空域失真；(3)计算单片段的时空域失真；(4)融合单片段的时空域失真得到单帧的失真，然后进一步融合得到全景视频的失真.

(3)基于学习的方法

类似于前一类方法，该类方法主要考虑用户观看全景视频过程中的局部视口感知特性，并通过学习的方式构建局部感知至全局质量的映射模型. Li等人[197]将预测的HM和EM嵌入到CNN结构中，设计了一个针对全景视频质量评价的CNN模型，其中HM和EM的预测借助了DHP[234]和SalGAN[235]视觉注意力模型. 该CNN模型将输入的全景视频直接分成块处理；然后使用预测的HM对相应的块采样，经过训练的CNN预测每个块的质量分数；最后利用预测的EM对这些块进行加权得到最终的质量得分[197]. 然而此类基于块的方法不可避免地会引入投影差异带来的失真，并且在实际观看中，人看到的是一个个视口而非图像块，因此一系列基于视口的方法被相继提出. Li等人[236-237]提出了一种基于视

1 https://jvet.hhi.fraunhofer.de/svn/svn_360Lib/

口的CNN方法(V-CNN),该方法主要分为两个阶段:第一阶段为视口预测与提取阶段;第二阶段为视口质量计算阶段.第一阶段采用了球形CNN模型[238]和柔性非最大抑制算法提取视口的位置;第二阶段采用Mini-DenseNet[239]预测视口的显著性及其相应权重,视频最终的质量分数由提取的多个视口质量融合而成.ProVQA[240]考虑到人对全景视频质量的感知是渐进的,在模型框架上设计了三个子网络,分别从像素、帧和视频中逐步学习球面感知质量、运动感知质量和多帧时间非感知质量.Chai和Shao[241]提出一个基于双流CNN的全景视频质量评价模型,该模型将全景视频通过CMP投影得到的六个等面积二维视频作为输入;然后使用一个双流CNN模型来提取帧内和帧间信息用于对质量失真部分建模,其中训练的双流CNN和3D卷积模型分别用于对空间和时间质量特征进行建模.Yang等人[242]提出联合球形CNN和非局部注意力模块的双目全景视频质量评价模型.该模型的输入为间隔取样的多对左右全景视频帧的差异图,核心为两组球形卷积操作和非局部注意力模块;随后接一个球形卷积操作和全连接网络,输出为质量分数.Meng等人[198]将Q-STAR方法[243]引入到全景视频质量评价中,提出融合帧分辨率、帧率和量化值的显著视口视频质量评价方法以及快速扫视区域质量评价方法,将显著视口区域和快速扫视区域质量融合得到全景视频质量.

4.3 小结

不同于一般的二维视频,屏幕内容视频、三维视频、合成视频和全景视频有着它们各自的特点,针对这些视频内容设计的模型需要考虑视频内容的特点,如屏幕内容视频中不同部分的失真对视觉感知的影响存在差异、三维视频感知中存在双目视觉现象、合成视频中的非均匀失真和相邻帧闪烁失真对质量的影响、全景视频感知中的局部视口感知特性.因此,研究者往往从视频特点出发,结合失真情况等设计相应的客观VQA模型.对于主观实验,实验方法主要沿用一般二维视频主观实验的方法,实验环境均是实验室.现有的其他内容视频质量主观数据集的数据量相对较少,视频时长较短.

5 未来研究方向

5.1 视频内容方面

(1)视听媒体QoE评价

大部分现有的VQA研究都是针对视频单一信号的,据我们所知,Min等人[244-246]较早地开展了视听媒体QoE研究,其中视听媒体包括视频和音频.Min等人[244]主要研究视频压缩降质和音频压缩降质对用户QoE的影响,开展实验时排除了时延、丢帧和不同步等因素造成的影响.该方向目前尚处于初始研究阶段,有较大的研究空间.同时,融合如文字描述等多媒体信息的VQA研究也是潜在的一个发展方向[247-249].

(2)流媒体质量评价

互联网技术的发展使得流媒体业务逐步占据人们日常的网络应用,视频点播、视频直播和视频会议等业务的流量已成为互联网主流.虽然网络基础设施和视频技术在不断更新换代,但终端用户对于高品质视频的需求也越来越高.针对流媒体的质量评价也被越来越多的研究学者所重视.通常流媒体的失真包含时域失真和空域失真[250],其中,时域失真主要由网络传输产生,如局部变形、频繁闪烁和马赛克等,而空域失真则主要由编解码方式产生,如振铃效应、模糊效应和方块效应等.这些视觉效果会直接导致用户的QoE显著降低.因此,需要能够准确有效地评估流媒体视频质量的方法.Duanmu等人[32,251-252]提出基于HTTP的动态自适应流(DASH)的视频传输质量评价方法,通过构建综合的流媒体视频质量主观数据集,研究不同因素影响下基于HVS的视频质量评价方法.尽管该研究方向已得到初步探索,但仍然存在许多挑战问题亟待解决.

(3)长视频质量评价

当前研究的对象可以称为“短视频”,大部分视频时长大概在10 s左右,鲜有研究考虑“长视频”(如1分钟)质量评价问题.长时间序列的质量评价问题可能对主观实验的方式和客观模型的设计均有较大的影响.在主观实验的设计方面,问题可能包括:受试者应该一次性看完还是分多次看完?对应的就是一个长视频的标签如何给定?在客观模型方面,问题可能包括:如何根据单一标签或者多阶段标签建模长时间序列降质等.

5.2 模型方面

(1)新式媒体质量评价

随着移动互联网的快速发展,新式媒体越来越多地出现在我们的日常生活中,如UGC视频[60]、游戏视频[253-254]、屏幕内容视频和夜间视频[255-256]等,它们的内容形式更加多样,并且它们的失真类型更加复杂,根据视频内容和失真的特点设计相应的VQA模型是主要的研究方向.

(2)融合新技术的模型设计

近年来视频处理技术层出不穷，当前研究人员较为关注的一项技术称为注意力机制，它主要用于解决CNN中长短距离依赖的问题. 该技术也被应用到了VQA模型的设计中，并取得了不错的性能[257-259]. 另外，少量的训练数据量是影响模型性能的一个重要因素，CS[260-261]是解决数据量不足的一个主流手段. 研究者们使用对比学习在训练过程中引入大量的无标记数据，进而获得较好的网络初始化参数，然后用于下游任务的微调. Chen等人[165]较早地将CS引入到了VQA模型的设计中，该模型与现有主流VQA模型相比取得了较大的性能提升. 期待更多的VQA研究可以与新近提出的技术结合，进一步推动这个领域的发展.

(3)兼顾性能和效率的模型设计

模型性能和效率的兼顾是视频处理和视频理解领域一个经典的问题，研究的目的是期望模型在一定计算量的情况下能够获得不错性能. 在VQA研究中，研究者们也做了一些尝试[25-132]，主要的方式就是抽取视频的部分帧作为模型的输入以减少计算量，该方式的出发点是视频序列中存在着大量的冗余帧. 虽然VQA模型的输入仅包含部分帧，但是依然可以获得不错的性能[25,132]. 该研究方向值得进一步探索.

6 总 结

本文对VQA研究领域进行了详细的综述. 首先介绍了VQA的基础知识，包括常见的主观质量评价方法、客观质量评价的分类及定义和现有的模型性能评价指标. 然后根据内容的不同分别介绍了视频质量主观数据集和客观模型. 对于主观数据集，介绍了它们的基本构成；对于客观模型，重点介绍了它们的设计原理. 最后，根据现有的发展趋势指出了VQA未来潜在的研究方向.

参 考 文 献

[1] Wang Z, Wu B, Wang W, et al. A survey of social relation understanding based on image and video information. Chinese Journal of Computers, 2021, 44(2): 1168-1199 (in Chinese)
(王正，吴斌，王文哲等. 基于图像和视频信息的社交关系理解研究综述. 计算机学报，2021，44(2)：1168-1199)

[2] Jia C, Ma H, Yang W, et al. A survey of social relation understanding based on image and video information. Journal of Image and Graphics, 2021, 26(06): 1179-1200 (in Chinese)
(贾川民，马海川，杨文瀚等. 视频处理与压缩技术. 中国图象图形学报，2021，26(6)：1179-1200)

[3] Zhao Y, Tian Y, Dang J, et al. Frontiers of transportation video structural analysis in the smart city. Journal of Image and Graphics, 2021, 26(06): 1227-1253 (in Chinese)
(赵耀，田永鸿，党建武等. 面向智慧城市的交通视频结构化分析前沿进展. 中国图象图形学报，2021，26(06)：1227-1253)

[4] Li X, Zhao B. Video distillation. SCIENCE CHINA Information Sciences, 2021, 51: 695-734 (in Chinese)
(李学龙，赵斌. 视频萃取. 中国科学：信息科学，2021，51：695-734)

[5] Wang Z, Sheikh H R, Bovik A C. Modern image quality assessment. San Rafael. California. USA: Morgan & Claypool, 2006

[6] Wang Z, Bovik A C. Objective video quality assessment, in The Handbook of Video Databases: Design and Applications. Boca Raton. Florida. USA: CRC Press, 2003

[7] Wang Z, Bovik A C. Reduced- and no-reference image quality assessment: The natural scene statistic model approach. IEEE Signal Processing Magazine, 2011, 28(6): 29-40

[8] Wang Z. Applications of objective image quality assessment methods. IEEE Signal Processing Magazine, 2011, 28(6): 137-142

[9] Fang Y, Zhu H, Ma K, et al. Perceptual evaluation for multi-exposure image fusion of dynamic scenes. IEEE Transactions on Image Processing, 2019, 29: 1127-1138

[10] Fang Y, Zeng Y, Jiang W, et al. Superpixel-based quality assessment of multi-exposure image fusion for both static and dynamic scenes. IEEE Transactions on Image Processing, 2021, 30: 2526-2537

[11] Fang Y, Fang Z, Yuan F, et al. Optimized multioperator image retargeting based on perceptual similarity measure. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2016, 47(11): 2956-2966

[12] Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing, 2004, 13(4): 600-612

[13] Laparra V, Berardino A, Balle J, et al. Perceptually optimized image rendering. Journal of the Optical Society of America A, 2017, 34(9): 1511-1525

[14] Ma K, Yeganeh H, Zeng K, et al. High dynamic range image compression by optimizing tone mapped image quality index. IEEE Transactions on Image Processing, 2015, 24(10): 3086-3097

[15] Ma K, Duanmu Z, Zhu H, et al. Deep guided learning for fast multi-exposure image fusion. IEEE Transactions on Image Processing, 2019, 29: 2808-2819

[16] Le C, Yan J, Fang Y, et al. Perceptually optimized deep high-dynamic-range image tone mapping//Proceedings of the IEEE International Conference on Virtual Reality and Visualization. Nanchang, China, 2021: 1-5

[17] Ding K, Ma K, Wang S, et al. Comparison of full-reference image quality models for optimization of image processing

systems. International Journal of Computer Vision, 2021, 129(4): 1258-1281

[18] Wang Z, Rehman A. Begin with the end in mind: A unified end-to-end quality-of-experience monitoring, optimization and management framework. SMPTE Motion Imaging Journal, 2019, 128(2): 1-8

[19] Fang Y, Zhu H, Zeng Y, et al. Perceptual quality assessment of smartphone photography//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual, 2020: 3677-3686

[20] Ma K, Duanmu Z, Wu Q, et al. Waterloo exploration database: New challenges for image quality assessment models. IEEE Transactions on Image Processing, 2016, 26(2): 1004-1016

[21] Wu J, Ma J, Liang F, et al. End-to-end blind image quality prediction with cascaded deep neural network. IEEE Transactions on Image Processing, 2020, 29: 7414-7426

[22] Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 2015, 115(3): 211-252

[23] Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA, 2014: 1725-1732

[24] Fabian C H, Victor E, Bernard G, et al. ActivityNet: A large-scale video benchmark for human activity understanding//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA, 2015: 961-970

[25] Yan J, Li J, Fang Y, et al. Subjective and objective quality of experience of free viewpoint videos. IEEE Transactions on Image Processing, 2022, 31: 3896-3907

[26] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks for action recognition in videos. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(11): 2740-2755

[27] Bhardwaj S, Srinivasan M, Khapra M M. Efficient video classification using fewer frames//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA, 2019: 354-363

[28] Wang L, Tong Z, Ji B, et al. TDN: Temporal difference networks for efficient action recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 1895-1904

[29] Feichtenhofer G. X3D: Expanding architectures for efficient video recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual, 2020: 203-213

[30] Wang S, Rehman A, Wang Z, et al. SSIM-motivated rate-distortion optimization for video coding. IEEE Transactions on Circuits and Systems for Video Technology, 2011, 22(4): 516-529

[31] Wang S, Rehman A, Wang Z, et al. Perceptual video coding based on SSIM-inspired divisive normalization. IEEE Transactions on Image Processing, 2012, 22(4): 1418-1429

[32] Duanmu Z, Ma K, Wang Z. Quality-of-experience for adaptive streaming videos: An expectation confirmation theory motivated approach. IEEE Transactions on Image Processing, 2018, 27(12): 6135-6146

[33] Varghese G, Wang Z. Video denoising based on a spatiotemporal Gaussian scale mixture model. IEEE Transactions on Circuits and Systems for Video Technology, 2010, 20(7): 1032-1040

[34] Yeh C, Lo K S, Lin W. Visual-quality guided global backlight dimming for video display on mobile devices. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 29(11): 3393-3403

[35] Zhang H, Zhang Y, Zhu L, et al. Deep learning-based perceptual video quality enhancement for 3D synthesized view. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(8): 5080-5094

[36] Liu J, Yang W, Yang S, et al. D3R-Net: Dynamic routing residue recurrent network for video rain removal. IEEE Transactions on Image Processing, 2018, 28(2): 699-712

[37] Yang W, Liu J, Feng J. Frame-consistent recurrent video deraining with dual-level flow//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA, 2019: 1661-1670

[38] Yang W, Tan R T, Wang S, et al. Self-learning video rain streak removal: When cyclic consistency meets temporal correspondence//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA, 2020: 1720-1729

[39] Lin W, Kuo C C J. Perceptual visual quality metrics: A survey. Journal of Visual Communication and Image Representation, 2011, 22(4): 297-312

[40] Athar S, Wang Z. A comprehensive performance evaluation of image quality assessment algorithms. IEEE Access, 2019, 7: 140030-140070

[41] Zhai G, Min X. Perceptual image quality assessment: A survey. SCIENCE CHINA Information Sciences, 2020, 63(11): 1-52

[42] Min X, Gu K, Zhao G, et al. Screen content quality assessment: Overview, benchmark, and beyond. ACM Computing Surveys, 2021, 54(9): 1-36

[43] Fang Y, Sui X, Yan J, et al. Progress in no-reference image quality assessment. Journal of Image and Graphics, 2021, 26(02): 0265-0286 (in Chinese)
(方玉明, 眭相杰, 鄢杰斌等. 无参考图像质量评价研究进展. 中国图象图形学报, 2021, 26(02): 0265-0286)

[44] Yan J, Fang Y, Liu X. Survey on image quality assessment from the perspective of distortion. Journal of Image and Graphics, 2022, 27(5): 1430-1466 (in Chinese)
(鄢杰斌, 方玉明, 刘学林. 图像质量评价综述-从失真的角度. 中国图象图形学报, 2022, 27(5): 1430-1466)

[45] Chen R, Yu Y, Shi D, et al. The review of image and video quality assessment methods. Journal of Image and Graphics, 2022, 27(5): 1410-1429 (in Chinese)
(程茹秋, 余烨, 石岱宗等. 图像与视频质量评价综述. 中国图象图形学报, 2022, 27(5): 1410-1429)

[46] Chikkerur S, Sundaram V, Reisslein M, et al. Objective video

quality assessment methods：A classification，review，and performance comparison. IEEE Transactions on Broadcasting，2011，57(2)：165-182

[47] Li D，Jiang T，Jiang M. Recent advances and challenges in video quality assessment. ZTE Communications，2019，17(1)：3-11

[48] Jiao L，Yang S，Liu F，et al. Seventy years beyond neural networks：Retrospect and prospect. Chinese Journal of Computers，2016，39(8)：1697-1916 (in Chinese)
(焦李成，杨淑媛，刘芳等. 神经网络七十年：回顾与展望. 计算机学报，2016，39(8)：1697-1916)

[49] Yuan F，Zhang L，Shi J，et al. Theories and applications of auto-encoder neural networks：A literature survey. Chinese Journal of Computers，2019，42(1)：203-230 (in Chinese)
(袁非牛，章琳，史劲亭等. 自编码神经网络理论及应用综述. 计算机学报，2019，42(1)：203-230)

[50] Li Y，Gao Y. Yan J，et al. Image inpainting methods based on deep neural networks：A review. Chinese Journal of Computers，2021，44(11)：2295-2316 (in Chinese)
(李月龙，高云，闫家良等. 基于深度神经网络的图像缺损修复方法综述. 计算机学报，2021，44(11)：2295-2316)

[51] Ji S，Du T，Deng S，et al. Robustness certification research on deep learning models：A survey. Chinese Journal of Computers，2022，45(1)：190-206 (in Chinese)
(纪守领，杜天宇，邓水光等. 深度学习模型鲁棒性研究综述. 计算机学报，2022，45(1)：190-206)

[52] Zhang W，Ma K，Zhai G，et al. Uncertainty-aware blind image quality assessment in the laboratory and wild. IEEE Transactions on Image Processing，2021，30：3474-3486

[53] Sinno Z，Bovik A C. Large-scale study of perceptual video quality. IEEE Transactions on Image Processing，2018，28(2)：612-627

[54] Ying Z，Mandal M，Ghadiyaram D，et al. Patch-VQ：'Patching up' the video quality problem//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual，2021：14019-14029

[55] Li J，Ling S，Wang J，et al. A probabilistic graphical model for analyzing the subjective visual quality assessment data from crowdsourcing//Proceedings of the ACM International Conference on Multimedia. Virtual，2020：3339-3347

[56] Li J，Ling S，Wang J，et al. GPM：A generic probabilistic model to recover annotator's behavior and ground truth labeling. arXiv preprint arXiv：2003.00475，2020

[57] ITU-T RECOMMENDATIONP.. Subjective video quality assessment methods for multimedia applications. International Telecommunication Union，1999

[58] RECOMMENDATIONITU-R. BT. 500-13. Methodology for the subjective assessment of the quality of television pictures. International Telecommunication Union，2012

[59] Vonikakis V，Subramanian R，Winkler S. Shaping datasets：Optimal data selection for specific target distributions across dimensions//Proceedings of the IEEE Conference on Image Processing. Phoenix，AZ，USA，2016：3753-3757

[60] Xu J，Li J，Zhou X，et al. Perceptual quality assessment of internet videos//Proceedings of the ACM International Conference on Multimedia. Virtual，2021：1248-1257

[61] Hahn F，Hosu V，Lin H，et al. KonVid-150k：A dataset for no-reference video quality assessment of videos in-the-wild. IEEE Access，2021，9：73139-72160

[62] Li G，Chen B，Zhu L，et al. PUGCQ：A large scale dataset for quality assessment of professional user-generated content//Proceedings of the ACM International Conference on Multimedia. Virtual，2021：3728-3736

[63] Pinson M，Barkowsky M，Callet P L. Selecting scenes for 2D and 3D subjective video quality tests. Journal on Image and Video Processing，2013，2013(1)：1-12

[64] Li J，Krasula L，Baveye Y，et al. AccAnn：A new subjective assessment methodology for measuring acceptability and annoyance of quality of experience. IEEE Transactions on Multimedia，2019，21(10)：2589-2606

[65] Simoncelli E P，Olshausen B. Natural image statistics and neural representation. Annual Review of Neuroscience，2001，24：1193-1216

[66] Wang Z，Bovik A C. A human visual system-based objective video distortion measurement system//Proceedings of the International Conference on Multimedia Processing and System，New York，USA，2000：1-4

[67] Wang Z，Shang X. Spatial pooling strategies for perceptual image quality assessment//Proceedings of the IEEE International Conference on Image Processing. Atlanta，GA，USA，2006：2945-2948

[68] Tu Z，Chen C，Chen L H，et al. A comparative evaluation of temporal pooling methods for blind video quality assessment//Proceedings of the IEEE International Conference on Image Processing. United Arab Emirates，2020：141-145

[69] Patrick L C，Christian V，Dominique B. A convolutional neural network approach for objective video quality assessment. IEEE Transactions on Neural Networks，2006，17(5)：1316-1327

[70] Liu W，Duanmu Z，Wang Z. End-to-end blind quality assessment of compressed videos using deep neural networks//Proceedings of the ACM International Conference on Multimedia. Seoul，Republic of Korea，2018：546-554

[71] VQEG. Final report from the video quality experts group on the validation of objective models of video quality assessment，2020

[72] Sheikh H R，Sabir M F，and Bovik A C. A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Transactions on Image Processing，2006，15(11)：3440-3451

[73] Ma K，Duanmu Z，Wu Q，et al. Waterloo exploration database：New challenges for image quality assessment models. IEEE Transactions on Image Processing，2016，26(2)：1004-1016

[74] Ma K，Duanmu Z，Wang Z，et al. Group maximum differentiation competition：Model comparison with few samples. IEEE Transactions on Pattern Analysis and Machine Intelligence，2020，42(4)：851-864

[75] Wang Z，Simoncelli E P. Maximum differentiation (MAD) competition：A methodology for comparing computational

models of perceptual quantities. Journal of Vision, 2008, 8(12): 1-13

[76] Wang H, Chen T, Wang Z, et al. I am going MAD: Maximum discrepancy competition for comparing classifiers adaptively// Proceedings of the IEEE Conference on Learning Representations, Addis Ababa, Ethiopia, 2020: 1-13

[77] Yan J, Zhong Y, Fang Y, et al. Exposing semantic segmentation failures via maximum discrepancy competition. International Journal of Computer Vision, 2021, 129(5): 1768-1786

[78] Wang H, Chen T, Wang Z, et al. Troubleshooting image segmentation models with human-in-the-loop. Machine Learning, 2023, 112: 1033-1051

[79] Wang Z, Wang H, Chen T, et al. Troubleshooting blind image quality models in the wild//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual, 2021: 16256-16265

[80] Wang Z, Ma K. Active fine-tuning from gMAD examples improves blind image quality assessment. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(9): 4577-4590

[81] Cao P, Wang Z, Ma K. Debiased subjective assessment of real-world image enhancement//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual, 2021: 711-721

[82] Simone F D, Naccari M, Tagliasacchi M, et al. Subjective assessment of H. 264/AVC video sequences transmitted over a noisy channel//Proceedings of the First International Workshop on Quality of Multimedia Experience. San Diego, California, USA, 2009: 204-209

[83] Seshadrinathan K, Soundararajan R, Bovik A C, et al. Study of subjective and objective quality assessment of video. IEEE Transactions on Image Processing, 2010, 19(6): 1427-1441

[84] Vu P V, Chandler D M. ViS3: An algorithm for video quality assessment via analysis of spatial and spatiotemporal slices. Journal of Electronic Imaging, 2014, 23(1): 013016

[85] Lin J Y, Song R, Wu C, et al. MCL-V: A streaming video quality assessment database. Journal of Visual Communication and Image Representation, 2015, 30: 1-9

[86] Wang H, Gan W, Hu S, et al. MCL-JCV: A JND-based H.264/AVC video quality assessment dataset//Proceedings of the IEEE International Conference on Image Processing. Phoenix, Arizona, USA, 2016: 1509-1513

[87] Nuutinen M, Virtanen T, Vaahteranoksa M, et al. CVD2014—A database for evaluating no-reference video quality assessment algorithms. IEEE Transactions on Image Processing, 2016, 25(7): 3073-3086

[88] Ghadiyaram D, Pan J, Bovik A C, et al. In-capture mobile video distortions: A study of subjective behavior and objective algorithms. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 28(9): 2061-2077

[89] Hosu V, Hahn F, Jenadeleh M, et al. The Konstanz natural video database//Proceedings of the Ninth International Conference on Quality of Multimedia Experience. Erfurt, Germany, 2017: 1-6

[90] Wang Y, Inguva S, Adsumilli B. YouTube UGC dataset for video compression research//Proceedings of the IEEE International Workshop on Multimedia Signal Processing. Kuala Lumpur, Malaysia, 2019: 1-5

[91] Li Y, Meng S, Zhang X, et al. User-generated video quality assessment: A subjective and objective study. IEEE Transactions on Multimedia, 2023, 25: 154-166

[92] Zhao K, Yuan K, Sun M, et al. Quality-aware pre-trained models for blind image quality assessment//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Vancouver, Canada 2023: 22302-22313

[93] Liu X, Song W, He Q, et al. Speeding up subjective video quality assessment via hybrid active learning. IEEE Transactions on Broadcasting, 2023, 69(1): 165-178

[94] Wang Z, Simoncelli E P. Reduced-reference image quality assessment using a wavelet-domain natural image statistic model//Proceedings of the 17th Annual Symposium on Electronic Imaging. San Jose, CA, USA, 2005: 149-159

[95] Sheikh H R, Bovik A C, Charrier C. No-reference quality assessment using natural scene statistics: JPEG2000. IEEE Transactions on Image Processing, 2005, 14(11): 1918-1927

[96] Moorthy A K, Bovik A C. Blind image quality assessment: From natural scene statistics to perceptual quality. IEEE Transactions on Image Processing, 2011, 20(12): 3350-3364

[97] Moorthy A K, Bovik A C. Blind image quality assessment: A natural scene statistics approach in the DCT domain. IEEE Transactions on Image Processing, 2012, 21(8): 3339-3352

[98] Mittal A, Moorth A K, Bovik A C. No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing, 2012, 21(12): 4695-4708

[99] Fang Y, Ma K, Wang Z, et al. No-reference quality assessment of contrast-distorted images based on natural scene statistics. IEEE Signal Processing Letters, 2014, 22(7): 818-842

[100] Fang Y, Yan J, Li L, et al. No reference quality assessment for screen content images with both local and global feature representation. IEEE Transactions on Image Processing, 2017, 27(4): 1600-1610

[101] Fang Y, Yan J, Du R, et al. Blind quality assessment for tone-mapped images by analysis of gradient and chromatic statistics. IEEE Transactions on Multimedia, 2020, 23: 955-966

[102] Saad M A, Bovik A C, Charrier C. Blind prediction of natural video quality. IEEE Transactions on Image Processing, 2014, 23(3): 1352-1365

[103] Li R, Zeng B, Liou M L. A new three-step search algorithm for block motion estimation. IEEE Transactions on Circuits and Systems for Video Technology, 1994, 4(4): 438-442

[104] Soundararajan R, Bovik A C. RRED Indices: Reduced reference entropic differencing for image quality assessment. IEEE Transactions on Image Processing, 2012, 21(2): 517-526

[105] Soundararajan R, Bovik A C. Video quality assessment by reduced reference spatio-temporal entropic differencing. IEEE Transactions on Circuits Systems and Video Technology,

2013，23(4)：684-694

[106] Bampis C G，Gupta P，Soundararajan R，et al. SpEED-QA：Spatial efficient entropic differencing for image and video quality. IEEE Signal Processing Letters，2017，24(9)：1333-1337

[107] Li X，Guo Q，Lu X. Spatiotemporal statistics for video quality assessment. IEEE Transactions on Image Processing，2016，25(7)：3329-3342

[108] Mittal A，Saad M A，Bovik A C. A completely blind video integrity oracle. IEEE Transactions on Image Processing，2016，25(1)：289-300.

[109] Dendi S V R，Channappayya S S. No-reference video quality assessment using natural spatiotemporal scene statistics. IEEE Transactions on Image Processing，2020，29：5612-5624

[110] Yu X，Birkbeck N. Wang Y，et al. Predicting the quality of compressed videos with pre-existing distortions. IEEE Transactions on Image Processing，2021，30：7511-7526

[111] Ebenezer J P，Shang Z. Wu Y，et al. ChipQA：No-reference video quality prediction via space-time chips. IEEE Transactions on Image Processing，2021，30：8059-8074

[112] Wang Z，Li Q. Information content weighting for perceptual image quality assessment. IEEE Transactions on Image Processing，2010，20(5)：1185-1198

[113] Fang Y，Yan J，Liu J，et al. Objective quality assessment of screen content images by uncertainty weighting. IEEE Transactions on Image Processing，2017，26(4)：2016-2027

[114] Wang Z，Simoncelli E P，Bovik A C. Multiscale structural similarity for image quality assessment//Proceedings of the Thirty-Seventh Asilomar Conference on Signals，Systems & Computers. Pacific Grove，CA，USA，2003：1398-1402

[115] Wang Z，Lu，L，Bovik A C. Video quality assessment based on structural distortion measurement. Signal Processing：Image Communication，2004，19(2)：121-132

[116] Pinson，M H，Wolf S. A new standardized method for objectively measuring video quality. IEEE Transactions on Broadcasting，2004，50(3)：312-322

[117] Li Q，Wang Z. Video quality assessment by incorporating a motion perception model//Proceedings of the IEEE International Conference on Image Processing. San Antonio，Texas，USA，2007：173-176

[118] Ninassi A，Le Meur O，Callet P L，et al. Considering temporal variations of spatial visual distortions in video quality assessment. IEEE Journal of Selected Topics in Signal Processing，2009，3(2)：253-265

[119] Seshadrinathan K，Bovik A C. Motion tuned spatio-temporal quality assessment of natural videos. IEEE Transactions on Image Processing，2009，19(2)：335-350

[120] Lu W，Li X，Gao X，et al. A video quality assessment metric based on human visual system. Cognitive Computation，2010，2(2)：120-131

[121] Sheikh H R，Bovik A C. Image information and visual quality. IEEE Transactions on Image Processing，2006，15(2)：430-444

[122] Li S，Zhang F，Ma L，et al. Image quality assessment by separately evaluating detail losses and additive impairments. IEEE Transactions on Multimedia，2011，13(5)：935-949

[123] Bampis C G，Li Z，Bovik A C. Spatiotemporal feature integration and model fusion for full reference video quality assessment. IEEE Transactions on Circuits and Systems for Video Technology，2018，29(8)：2256-2270

[124] Rehman A，Zeng K，Wang Z. Display device-adapted video quality of experience assessment//Proceedings of the Human Vision and Electronic Imaging XX. San Francisco，California，USA，2015：939406

[125] Zeng K，Wang Z. 3D-SSIM for video quality assessment//Proceedings of the IEEE International Conference on Image Processing. Orlando，FL，USA，2012：621-624

[126] Park J，Seshadrinathan K，Lee S，et al. Video quality pooling adaptive to perceptual distortion severity. IEEE Transactions on Image Processing，2013，22(2)：610-620

[127] Larson E C，Chandler D M. Most apparent distortion：Full-reference image quality assessment and the role of strategy. Journal of Electronic Imaging，2010，19(1)：011006

[128] Manasa L，Channappayya S S. An optical flow-based full reference video quality assessment algorithm. IEEE Transactions on Image Processing，2016，25(6)：2480-2492

[129] Banitalebi-Dehkordi M，Ebrahimi-Moghadam A，Khademi M，et al. No-reference video quality assessment based on visual memory modeling. IEEE Transactions on Broadcasting，2019，66(3)：676-689

[130] Korhonen J. Two-level approach for no-reference consumer video quality assessment. IEEE Transactions on Image Processing，2019，28(12)：5923-5938

[131] Wu J，Liu Y，Dong W，et al. Quality assessment for video with degradation along salient trajectories. IEEE Transactions on Multimedia，2019，21(11)：2738-2749

[132] Tu Z，Wang Y. Birkbeck N，et al. UGC-VQA：Benchmarking blind video quality assessment for user generated content. IEEE Transactions on Image Processing，2021，30：4449-4464

[133] Kancharla P，Channappayya S. S. Completely blind quality assessment of user generated video content. IEEE Transactions on Image Processing，2022，31：263-274

[134] Gao F，Gao X. Active feature learning and its application in blind image quality assessment. Chinese Journal of Computers，2014，37(10)：2227-2234 (in Chinese)
(高飞，高新波. 主动特征学习及其在盲图像质量评价中的应用. 计算机学报，2014，37(10)：2227-2234)

[135] Xu J，Ye P，Liu Y. No-reference video quality assessment via feature learning//Proceedings of the IEEE International Conference on Image Processing. Paris，France，2014：491-495

[136] Ye P，Kumar J，Kang L，et al. Unsupervised feature learning framework for no-reference image quality assessment//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. RI，USA，2012：1098-1105

[137] Xue W，Zhang L，Mou X，et al. Gradient magnitude similarity deviation：A highly efficient perceptual image quality index. IEEE Transactions on Image Processing，2013，23(2)：684-695

[138] Wang C, Su L, Zhang W, et al. No reference video quality assessment based on 3D convolutional neural network. Journal of Software, 2016, 27(2): 103-112 (in Chinese)
(王春峰, 苏荔, 张维刚等. 基于3D卷积神经网络的无参考视频质量评价. 软件学报, 2016, 27(2): 103-112)

[139] Wang C, Su L, Huang Q, et al. Spatio-temporal-fused no-reference video quality assessment based on convolutional neural network. Journal of University of Chinese Academy of Sciences, 2018, 35(4): 544-549 (in Chinese)
(王春峰, 苏荔, 黄庆明等. 基于卷积神经网络的时空融合的无参考视频质量评价方法. 中国科学院大学学报, 2018, 35(4): 544-549)

[140] Kang L, Ye P, Li Y, et al. Convolutional neural networks for no-reference image quality assessment//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA, 2014: 1733-1740

[141] Kim W, Kim J, Ahn S, et al. Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 219-234

[142] Zhang Y, Gao X, He L, et al. Blind video quality assessment with weakly supervised learning and resampling strategy. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 29(8): 2244-2255

[143] Li D, Jiang T, Jiang M. Quality assessment of in-the-wild videos//Proceedings of the ACM International Conference on Multimedia. Nice, France, 2019: 2351-2359

[144] Chen P, Li L, Ma L, et al. RIRNet: Recurrent-in-current network for video quality assessment//Proceedings of the Proceedings of the ACM International Conference on Multimedia. Seattle, WA, USA, 2020: 834-842

[145] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916

[146] Ying Z, Niu H, Gupta P, et al. From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual, 2020: 3572-3582

[147] Hara K, Kataoka H, Satoh Y. Learning spatio-temporal features with 3D residual networks for action recognition//Proceedings of the IEEE Conference on Computer Vision Workshops. Venice, Italy, 2017: 3154-3160

[148] Ismail F H, Lucas B, Forestier G, et al. InceptionTime: Finding AlexNet for time series classification. Data Mining and Knowledge Discovery, 2020, 34(6): 1936-1962

[149] Yi F, Chen M, Sun W, et al. Attention based network for no-reference UGC video quality assessment//Proceedings of the IEEE International Conference on Image Processing. Anchorage, Alaska, USA, 2021: 1414-1418

[150] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556, 2014

[151] Chen J, Wang H, Xu M, et al. Deep neural networks for end-to-end spatiotemporal video quality prediction and aggregation//Proceedings of the IEEE International Conference on Multimedia and Expo. Virtual, 2021: 1-6

[152] Zhu H, Chen B, Zhu L, et al. Learning spatiotemporal interactions for user-generated video quality assessment. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(3): 1031-1042

[153] Wu H, Chen C, Hou J, et al. Fast-VQA: Efficient end-to-end video quality assessment with fragment sampling//Proceedings of the European Conference on Computer Vision. Tel AVIV, Israel, 2022: 538-554

[154] Liu Y, Wu J, Li A, et al. Video quality assessment with serial dependence modeling. IEEE Transactions on Multimedia, 2021, 24: 3754-3768

[155] Wu W, Li Q, Chen Z, et al. Semantic information oriented no-reference video quality assessment. IEEE Signal Processing Letters, 2021, 28: 204-208

[156] Shen W, Zhou M, Liao X, et al. An end-to-end no-reference video quality assessment method with hierarchical spatiotemporal feature representation. IEEE Transactions on Broadcasting, 2022, 68(3): 651-660

[157] Li Z, Yang L. DCVQE: A hierarchical transformer for video quality assessment//Proceedings of the Asian Conference on Computer Vision. Macau SAR, China, 2022: 2562-2579

[158] You J, Lin Y. Efficient transformer with locally shared attention for video quality assessment//Proceedings of the IEEE International Conference on Image Processing. Bordeaux, France, 2022: 356-360

[159] Xing F, Wang Y, Wang H, et al. StarVQA: Space-time attention for video quality assessment//Proceedings of the IEEE International Conference on Image Processing. Bordeaux, France, 2022: 2326-2330

[160] Guan X, Li F, Zhang Y, et al. End-to-end blind video quality assessment based on visual and memory attention modeling. IEEE Transactions on Multimedia, 2022, to appear

[161] Ma K, Liu W, Wang Z. End-to-end image quality assessment using deep neural networks. IEEE Transactions on Image Processing, 2017, 27(3): 1202-1213

[162] Wang Y, Ke J, Talebi H, et al. Rich features for perceptual quality assessment of UGC videos//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 13435-13444

[163] Liu Y, Wu J, Li A, et al. Spatiotemporal representation learning for blind video quality assessment. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 23(2): 684-695

[164] Liu Y, Wu J, Li A, et al. No-reference video quality assessment with heterogeneous knowledge ensemble//Proceedings of the ACM International Conference on Multimedia. Chengdu, China, 2021: 4174-4182

[165] Chen P, Li L, Wu J, et al. Contrastive self-supervised pre-training for video quality assessment. IEEE Transactions on Image Processing, 2021, 31: 458-471

[166] Li B, Zhang W, Tian M, et al. Blindly assess quality of in-the-

wild videos via quality-aware pre-training and motion perception. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(9): 5944-5958

[167] Mitra S, Soundararajan R. Multiview contrastive learning for completely blind video quality assessment of user generated content//Proceedings of the ACM International Conference on Multimedia. Lisbon, Portugal, 2022: 1914-1924

[168] Jiang S, Sang Q, Hu Z, et al. Self-supervised representation learning for video quality assessment. IEEE Transactions on Broadcasting, 2023, 69(1): 118-129

[169] Li D, Jiang T, Jiang M. Unified quality assessment of in-the-wild videos with mixed datasets training. International Journal of Computer Vision, 2021, 129(4): 1238-1257

[170] Chen P, Li L, Wu J, et al. Unsupervised curriculum domain adaptation for no-reference video quality assessment// Proceedings of the IEEE Conference on Computer Vision. Virtual, 2021: 5178-5187

[171] Chen B, Zhu L, Li Guo, et al. Learning generalized spatial-temporal deep feature representation for no-reference video quality assessment. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(4): 1903-1916

[172] Xian W, Zhou M, Fang B, et al. Spatiotemporal feature hierarchy-based blind prediction of natural video quality via transfer learning. IEEE Transactions on Broadcasting, 2022, 69(1): 130-143

[173] Chen P, Li L, Li H, et al. Dynamic expert-knowledge ensemble for generalizable video quality assessment. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(6): 2577-2589

[174] Yan J. The Research on Quality Assessment of Stereo Visual Signals [M. S. thesis]. Jiangxi University of Finance and Economics, Nanchang, 2018 (in Chinese)
(鄢杰斌. 基于双目视觉感知的无参考三维图像视觉质量评价[硕士学位论文]. 江西财经大学, 南昌, 2018)

[175] Fang Y, Yan J, Liu J, et al. Objective quality assessment of screen content images by uncertainty weighting. IEEE Transactions on Image Processing, 2017, 26(4): 2016-2027

[176] Fang Y, Yan J, Li L, et al. No reference quality assessment for screen content images with both local and global feature representation. IEEE Transactions on Image Processing, 2017, 27(4): 1600-1610

[177] Zhou Y. Study on the Objective Quality Assessment Metrics of Virtual View Synthesis [Ph.D. dissertation]. China University of Mining and Technology, Xuzhou, 2019 (in Chinese)
(周玉. 面向虚拟视觉合成的客观质量评价方法研究[博士学位论文]. 中国矿业大学, 徐州, 2019)

[178] Yan J, Fang Y, Ru R, et al. No reference quality assessment for 3D synthesized views by local structure variation and global naturalness change. IEEE Transactions on Image Processing, 2020, 29: 7443-7453

[179] Cheng S, Zeng H, Chen J, et al. Screen content video quality assessment: Subjective and objective study. IEEE Transactions on Image Processing, 2020, 29: 8636-8651

[180] Li T, Min X, Zhao H, et al. Subjective and objective quality assessment of compressed screen content videos. IEEE Transactions on Broadcasting, 2020, 67(2): 438-449

[181] Chen M, Kwon D, Bovik A C. Study of subject agreement on stereoscopic video quality//Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation. Santa Fe, NM, USA, 2012: 173-176

[182] Urvoy M, Barkowsky M, Cousseau R, et al. NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences//Proceedings of the International Workshop on Quality of Multimedia Experience. Yarra Valley, Australia, 2012: 109-114

[183] Silva V, Arachchi H, Ekmekcioglu E, et al. Toward an impairment metric for stereoscopic video: A full-reference video quality metric to assess compressed stereoscopic video. IEEE Transactions on Image Processing, 2013, 22(9): 3392-3404

[184] Wang J, Wang S, Wang Z. Asymmetrically compressed stereoscopic 3D videos: Quality assessment and rate-distortion performance evaluation. IEEE Transactions on Image Processing, 2017, 26(3): 1330-1343

[185] Appina B, Dendi S V R, Manasa K, et al. Study of subjective quality and objective blind quality prediction of stereoscopic videos. IEEE Transactions on Image Processing, 2019, 28(10): 5027-5040

[186] Bosc E, Pepion R, Callet P L, et al. Perceived quality of DIBR-based synthesized views//Proceedings of the Applications of Digital Image Processing XXXIV. San Diego, California, USA, 2011: 1-9

[187] Bosc E, Hanhart P, Callet P L, et al. A quality assessment protocol for free-viewpoint video sequences synthesized from decompressed depth data//Proceedings of the International Workshop of Quality of Multimedia Experience. Klagenfurt am Worthersee, Austria, 2013: 1-9

[188] Liu X, Zhang Y, Hu S, et al. Subjective and objective video quality assessment of 3D synthesized with texture/depth compression distortion. IEEE Transactions on Image Processing, 2022, 24(12): 4847-4861

[189] Ling S, Gutierrez J, Gu K, et al. Prediction of the influence of navigation scan-path on perceived quality of free-viewpoint videos. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2019, 9(1): 204-216

[190] Singla A, Fremerey S, Robitza W, et al. Comparison of subjective quality evaluation for HEVC encoded omnidirectional videos at different bit-rates for UHD and FHD resolution// Proceedings of the on Thematic Workshops of ACM Multimedia. Mountain View, CA, USA, 2017: 511-519

[191] Curcio I D D, Toukomaa H, Naik D. Bandwidth reduction of omnidirectional viewport-dependent video streaming via subjective quality assessment//Proceedings of the 2nd International Workshop on Multimedia Alternate Realities. Mountain View, CA, USA, 2017: 9-14

[192] Tran H T T, Ngoc N P, Bui C M, et al. An evaluation of quality metrics for 360 videos//Proceedings of the 2017 Ninth International Conference on Ubiquitous and Future Networks.

Milan, Italy, 2017: 7-11

[193] Duan H, Zhai G, Yang X, et al. IVQAD 2017: An immersive video quality assessment database//Proceedings of the International Conference on Systems, Signals and Image Processing. Sofia, Bulgaria, 2017: 1-5

[194] Zhang B, Zhao J, Yang S, et al. Subjective and objective quality assessment of panoramic videos in virtual reality environments//Proceedings of the IEEE International Conference on Multimedia & Expo Workshops. Hong Kong, China, 2017: 163-168

[195] Zhang Y, Wang Y, Liu F, et al. Subjective panoramic video quality assessment database for coding applications. IEEE Transactions on Broadcasting, 2018, 64(2): 461-473

[196] Lopes F, Ascenso J, Rodrigues A, et al. Subjective and objective quality assessment of omnidirectional video//Proceedings of the Applications of Digital Image Processing XLI. San Diego, CA, USA, 2018: 249-265

[197] Li C, Xu M, Du X, et al. Bridge the gap between VQA and human behavior on omnidirectional video: A large-scale dataset and a deep learning model//Proceedings of the 26th ACM International Conference on Multimedia. New York, NY, USA, 2018: 932-940

[198] Meng Y, Ma Z. Viewport-based omnidirectional video quality assessment: database, modeling and inference. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(1): 120-134

[199] Zou W, Zhang W, Yang F. Modeling the perceptual quality for viewport-adaptive omnidirectional video streaming considering dynamic quality boundary artifact. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(11): 4241-4254

[200] Mahmoudpour S, Schelkens P. Omnidirectional video quality index accounting for judder. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(1): 61-75

[201] Li T, Min X, Zhu W, et al. No-reference screen content video quality assessment. Displays, 2021, 69: 102030

[202] Zeng H, Huang H, Hou J, et al. Screen content video quality assessment model using hybrid spatiotemporal features. IEEE Transactions on Image Processing, 2022, 31: 6175-6187

[203] Yasakethu S, Hewage C, Fernando W, et al. Quality analysis for 3D video using 2D video quality models. IEEE Transactions on Consumer Electronics, 2008, 54(4): 1969-1976

[204] Wang J, Rehman A, Zeng K. Quality prediction of asymmetrically distorted stereoscopic 3D images. IEEE Transactions on Image Processing, 2015, 24(11): 3400-3414

[205] Wang J, Wang S, Ma K, et al. Perceptual depth quality in distorted stereoscopic images. IEEE Transactions on Image Processing, 2016, 26(3): 1202-1215

[206] Appina B, Channappayya S S. Full-reference 3D video quality assessment using scene component statistical dependencies. IEEE Signal Processing Letters, 2018, 25(6): 823-827

[207] Jiang G, Liu S, Yu M, et al. No reference stereo video quality assessment based on motion feature in tensor decomposition domain. Journal of Visual Communication and Image Representation, 2018, 50: 247-262

[208] Chen Z, Zhou W, Li W. Blind stereoscopic video quality assessment: From depth perception to overall experience. IEEE Transactions on Image Processing, 2017, 27(2): 721-734

[209] Yang J, Zhao Y, Jiang B, et al. No-reference quality evaluation of stereoscopic video based on spatio-temporal texture. IEEE Transactions on Multimedia, 2019, 22(10): 2635-2644

[210] Appina B, Dendi S V R, Manasa K, et al. Study of subjective quality and objective blind quality prediction of stereoscopic videos. IEEE Transactions on Image Processing, 2019, 28(10): 5027-5040

[211] Qi F, Zhao D, Fan X, et al. Stereoscopic video quality assessment based on visual attention and just-noticeable difference models. Signal, Image and Video Processing, 2016, 10(4): 737-744

[212] Yang J, Zhao Y, Jiang B, et al. No-reference quality assessment of stereoscopic videos with inter-frame cross on a content-rich database. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(10): 3608-3623

[213] Fang Y, Sui X, Wang J, et al. Perceptual quality assessment for asymmetrically distorted stereoscopic video by temporal binocular rivalry. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(8): 3010-3024

[214] Fang Y, Sui X, Yan J, et al. Asymmetrically distorted 3D video quality assessment: From the motion variation to perceived quality. Signal Processing, 2021, 183: 108031

[215] Zhou W, Chen Z, Li W. Stereoscopic video quality prediction based on end-to-end dual stream deep neural networks//Proceedings of the Pacific Rim Conference on Multimedia. Hefei, China, 2018: 482-492

[216] Feng Y, Li S, Chang Y. Multi-scale feature-guided stereoscopic video quality assessment based on 3D convolutional neural network//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto, Ontario, Canada, 2018: 2095-2099

[217] Yang J, Zhu Y, Ma C, et al. Stereoscopic video quality assessment based in 3D convolutional neural networks. Neurocomputing, 2018, 309: 83-93

[218] Sun C, Liu X, Yang W, et al. An efficient quality metric for DIBR-based 3D video//Proceedings of the International Conference on High Performance Computing and Communication & 9th International Conference on Embedded Software and Systems, Washington, DC, USA, 2012: 1391-1394

[219] Kim H G, Ro Y M. Measurement of critical temporal inconsistency for quality assessment of synthesized videos//Proceedings of the IEEE International Conference on Image Processing, Phoenix, Arizona, 2016: 1027-1031

[220] Huang Y, Zhou Y, Hu B, et al. DIBR-synthesized video quality assessment by measuring geometric distortion and spatiotemporal inconsistency. Electronics Letters, 2020, 56(24): 1314-1317

[221] Zhang Y, Zhang H, Yu M, et al. Sparse representation-based video quality assessment for synthesized 3D videos. IEEE Transactions on Image Processing, 2019, 29: 509-524

[222] Stankovic D S, Callet P L, Battisti F, et al. Free viewpoint video quality assessment based on morphological multiscale metrics//Proceedings of the International Conference on Quality of Multimedia Experience, Lisbon, Portugal, 2016: 1-6

[223] Ling S, Li J, Che Z, et al. Quality assessment of free-viewpoint videos by quantifying the elastic changes of multi-scale motion trajectories. IEEE Transactions on Image Processing, 2021, 30: 517-531

[224] Zhou Y, Li L, Wang S, et al. No-reference quality assessment of DIBR-synthesized videos by measuring temporal flickering. Journal of Visual Communication and Image Representation, 2018, 55: 30-39

[225] Wang G, Wang Z, Gu K, et al. Reference-free DIBR-synthesized video quality metric in spatial and temporal domains. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(3): 1119-1132

[226] Ling S, Li J, Che Z, et al. Re-visiting discriminator for blind free-viewpoint image quality assessment. IEEE Transactions on Multimedia, 2020, 23: 4245-4258

[227] Zakharchenko V, Alshina E, Singh A, et al. AhG8: Suggested testing procedure for 360-degree video//Joint Video Exploration Team of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-D0027, 4th Meeting. 2016

[228] Sun Y, Lu A, Yu L. Weighted-to-spherically-uniform quality evaluation for omnidirectional video. IEEE Signal Processing Letters, 2017, 24(9): 1408-1412

[229] Yu M, Lakshman H, Girod B. A framework to evaluate omnidirectional video coding schemes//Proceedings of the IEEE International Symposium on Mixed and Augmented Reality. Fukuoka, Japan, 2015: 31-36

[230] Chen S, Zhang Y, Li Y, et al. Spherical structural similarity index for objective omnidirectional video quality assessment//Proceedings of the IEEE International Conference on Multimedia and Expo. San Diego, USA, 2018: 1-6

[231] Xu M, Li C, Chen Z, et al. Assessing visual quality of omnidirectional videos. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 29(12): 3516-3530

[232] Jiang Z, Xu Y, Sun J, et al. Tile-based panoramic video quality assessment. IEEE Transactions on Broadcasting, 2022, 68(2): 530-544

[233] Gao P, Zhang P, Smolic A. Quality assessment for omnidirectional video: A spatio-temporal distortion modeling approach. IEEE Transactions on Multimedia, 2022, 24: 1-16

[234] Xu M, Song Y, Wang J, et al. Predicting head movement in panoramic video: A deep reinforcement learning approach. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(11): 2693-2708

[235] Pan J, Ferrer C C, McGuinness K, et al. SalGAN: Visual saliency prediction with generative adversarial networks. arXiv preprint arXiv:1701.01081, 2017

[236] Li C, Xu M, Jiang L, et al. Viewport proposal CNN for 360° video quality assessment//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA, 2019: 10169-10178

[237] Xu M, Jiang L, Li C, et al. Viewport-based CNN: A multi-task approach for assessing 360° video quality. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(4): 2198-2215

[238] Cohen T S, Geiger M, Köhler J, et al. Spherical CNNs//Proceedings of the International Conference on Learning Representations. Vancouver, BC, Canada, 2018: 1-15

[239] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, Hawaii, USA, 2017: 4700-4708

[240] Yang L, Xu M, Li S, et al. Blind VQA on 360° video via progressively learning from pixels, frames and video. IEEE Transactions on Image Processing, 2022, 32: 128-143

[241] Chai X, Shao F. Blind quality assessment of omnidirectional videos using spatio-temporal convolutional neural networks. Optik, 2021, 226: 165887

[242] Yang J, Liu T, Jiang B, et al. Panoramic video quality assessment based on non-local spherical CNN. IEEE Transactions on Multimedia, 2021, 23: 797-809

[243] Qu Y, Xue Y, Wang Y. Q-STAR: A perceptual video quality model considering impact of spatial, temporal, and amplitude resolutions. IEEE Transactions on Image Processing, 2014, 23(6): 2473-2486

[244] Min X, Zhai G, Zhou J, et al. Study of subjective and objective quality assessment of audio-visual signals. IEEE Transactions on Image Processing, 2020, 29: 6054-6068

[245] Min X, Zhai G, Zhou J, et al. A multimodal saliency model for videos with high audio-visual correspondence. IEEE Transactions on Image Processing, 2020, 29: 3805-3819

[246] Yao S, Min X, Zhai G. Deep audio-visual fusion neural network for saliency estimation//Proceedings of the IEEE International Conference on Image Processing. Anchorage, Alaska, USA, 2021: 1604-1608

[247] Yang W, Wu J, Tian S, et al. Fine-grained image quality caption with hierarchical semantics degradation. IEEE Transactions on Image Processing, 2022, 31: 3578-3950

[248] Jiang W, Zhu M, Fang Y, et al. Visual cluster grounding for image captioning. IEEE Transactions on Image Processing, 2022, 31: 3920-3934

[249] Wan B, Jiang W, Fang Y. Informative attention supervision for grounded video description//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore, 2022: 1955-1959

[250] Stefan W, Ruth C. Video quality evaluation for internet streaming applications//Proceedings of the Human Vision and Electronic Imaging, 2003: 104-115

[251] Duanmu Z, Zeng K, Ma K, et al. A quality-of-experience index for streaming video. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(1): 154-166

[252] Duanmu Z, Rehman A, Wang Z. A quality-of-experience database for adaptive video streaming. IEEE Transactions on Broadcasting, 2018, 64(2): 474-487

[253] Yu X, Tu Z, Ying Z, et al. Subjective quality assessment of

user-generated content gaming videos//Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Hawaii, USA, 2021: 2112-2120

[254] Wen S, Ling S, Wang J, et al. Subjective and objective quality assessment of mobile gaming video//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore, 2022: 1810-1814

[255] Da P, Song G, Shi P, et al. Perceptual quality assessment of nighttime video. Displays, 2021, 70: 102092

[256] Guan X, Li F, Huang Z, et al. Study of subjective and objective quality assessment of night-time videos. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(10): 6627-6641

[257] You J. Long short-term convolutional transformer for no-reference video quality assessment//Proceedings of the ACM International Conference on Multimedia. Chengdu, China, 2021: 2112-2120

[258] You J, Zhang Z. Visual mechanisms inspired efficient transformers for image and video quality assessment. arXiv preprint arXiv:2203.14557, 2022

[259] Wu H, Chen C, Liao L, et al. DisCoVQA: Temporal distortion-content transformers for video quality assessment. IEEE Transactions on Circuits and Systems for Video Technology, 2023, Early Access

[260] Chen P, Li L, Wu Q, et al. SPIQ: A self-supervised pre-trained model for image quality assessment. IEEE Signal Processing Letters, 2022, 29: 513-517

[261] Madhusudana P C, Birkbeck N, Wang Y, et al. Image quality assessment using contrastive learning. IEEE Transactions on Image Processing, 2022, 31: 4149-41761

YAN Jie-Bin, Ph. D., lecturer. His research interests include visual quality assessment and computer vision.

FANG Yu-Ming, Ph. D., professor, Ph. D. supervisor. His research interests include multimedia signal processing and visual quality assessment.

LIU Xue-Lin, Ph. D. candidate. His research interest is visual quality assessment.

YAO Yi-Ru, M. S. candidate. Her research interest is visual quality assessment.

SUI Xiang-Jie, Ph. D. candidate. His research interest is visual quality assessment.

Background

Video quality assessment (VQA) is a basic problem in video processing and video understanding field, and its objective is to accurately predict the visual quality of videos automatically. Objective VQA models can be used to screen out those low-quality videos from massive videos with varied quality, which can largely reduce storage pressure. Besides, objective VQA modes can serve as benchmarks for evaluating other video processing algorithms, such as video coding, video enhancement, video transmission, etc. Since video quality degradation exists in every stage from video production to video consumption, this research topic has gained more and more attention from both academic and industry.

This paper reviews the studies regarding VQA published nearly in the past two decades, as well as the basic knowledge about subjective experiments and the design philosophy of objective models. More specifically, we introduce the composition of subjective video quality databases and the details about objective models. For a clear representation, we describe each part according to the video content, including general 2D video, screen content video, stereoscopic video, synthesized video and omnidirectional video.

This paper was supported partly by the National Natural Science Foundation of China (62132006), partly by the Natural Science Foundation of Jiangxi Province of China (20224BAB212012), partly by the project funded by China Postdoctoral Science Foundation (2022M721417), and partly by the Project of the Education Department of Jiangxi Province of China (GJJ2200524). VQA is one of the main research topics of our team led by Prof. Yuming Fang (He is a recipient of the National Science Fund for Excellent Young Scholars of China) in Jiangxi University of Finance and Economics. In the past few years, we have published a few relevant papers in the visual quality assessment topic.