doi:10.3969/j.issn.1001-4616.2025.03.012

# 考虑背景失真的无参考视频质量评价方法

朱文佳[1,2],张　婷[2],程茹秋[2],余　烨[2]

(1.安徽百诚慧通科技股份有限公司,安徽 合肥 230001)
(2.合肥工业大学计算机与信息学院,安徽 合肥 230000)

[**摘要**]　现实场景下拍摄的视频由于存在各种未知失真类型、缺少参考视频,对此类视频的质量评价是一个十分具有挑战性的任务. 近年来,研究人员将人类视觉系统的先验知识融合在质量评价任务中. 在此基础上,提出一种考虑背景失真的无参考视频质量评价方法. 该方法在考虑视频内容的同时,显著增强了对视频背景中信息丢失问题的敏感度,在特征提取阶段充分考虑背景特征的提取;随后,通过引入结合门控机制的通道挖掘技术,高效整合高低维特征,使特征通道更加精准地聚焦于背景失真细节;最终,利用时序建模模块构建特征的时间维度模型,并通过线性回归方法生成视频质量的客观量化评分. 使用 SROCC(spearman rank order correlation coefficient)、PLCC(pearson linear correlation coefficient)和 RMSE(root mean squared error)等评价指标在公开数据集 KoNViD-1k、LIVE-Qualcomm 和 CVD2014 开展实验,结果表明该方法不仅与人类主观感知具有高度相关性,且预测误差较小,有效提升了视频质量评估的准确性和可靠性,能够更贴近地模拟人类对视频质量的直观评价.

[**关键词**]　视频质量评价,无参考,背景失真,通道挖掘机制,时序建模
[**中图分类号**]TP391.41　[**文献标志码**]A　[**文章编号**]1001-4616(2025)03-0102-10

# No-Reference Video Quality Assessment Method Based on Background Distortions

Zhu Wenjia[1,2],Zhang Ting[2],Cheng Ruqiu[2],Yu Ye[2]

(1.Anhui Baichenghuitong Technology Co.,Ltd.,Hefei 230001,China)
(2.School of Computer and Information,Hefei University of Technology,Hefei 230000,China)

**Abstract**:Due to the existence of various unknown distortion types and the lack of reference videos,it is a very challenging task to evaluate the quality of such videos. In recent years,researchers have integrated the transcendental knowledge of human visual system into the task of quality evaluation. On this basis,a non-reference video quality assessment(VQA) method considering background distortion is proposed. This method not only considers the video content,but also significantly enhances the sensitivity to the information loss in the video background,and fully considers the extraction of background features in the feature extraction stage. Then,by introducing the channel mining technology combined with gating mechanism,the high and low dimensional features are efficiently integrated,so that the feature channels can focus on the details of background distortion more accurately. Finally,the time dimension model of features is constructed by using the time series modeling module,and the objective quantitative score of video quality is generated by linear regression method. The evaluation indexes such as SROC(Spearman rank order correlation coefficient),PLCC(Pearson linear correlation coefficient) and RMSE(root mean squared error) are used to carried out experiments on KoNViD-1k,LIVE-Qualcomm and CVD2014 datasets. The results show that this method not only has a high correlation with human subjective perception,but also has a small prediction error,which effectively improves the accuracy and reliability of VQA and can more closely simulate human visual evaluation of video quality.

**Key words**:video quality assessment(VQA),no-reference,background distortion,channel mining mechanism,time series modeling

**收稿日期**:2024-07-22.
**基金项目**:安徽省自然科学基金资助项目(2308085MF216)、国家自然科学基金资助项目(62372153).
**通讯作者**:朱文佳,硕士,工程师,研究方向:计算机视觉与智能交通;余烨,博士,副教授,研究方向:计算机视觉、低光图像增强及视频质量评价等. E-mail:yuye@ hfut.edu.cn

图像/视频质量评价(Image/Video Quality Accessment,IQA/VQA)作为计算机视觉的基本任务之一,在视频监控、工业检测及医疗影像等领域具有十分广泛的应用,具有重要的应用价值.

多项主观评估实验[1-3]显示,视频内容在 VQA 中占据核心地位,不同内容在相同观看条件下会引发观众显著的主观感受差异. 因此,当前主流的 VQA 技术[4-6]普遍采用基于 ImageNet 等大规模数据集预训练的卷积神经网络(convolutional neural networks,CNN)模型,以提取目标的语义特征,通过特征提取和分析,实现对视频质量的有效评估. 在聚焦拍摄目标的场景下,此类方法尤为有效,如图 1(a)所示,目标信息的丰富程度直接决定了视频的视觉吸引力和质量感知.

然而,在特定场景如长时间监控录像中,视频质量评估的维度变得更为复杂. 静止目标及设备故障引入的失真,使得背景信息在评估中的重要性凸显,如图 1(b)所示. 这表明在全面评价视频质量时,需综合考虑目标内容与背景环境的共同影响.



(a) 突出前景信息　　(b) 视频背景影响较大

**图 1　生活场景中拍摄的视频截图**

**Fig. 1　Video screenshots taken in ordinary life**

长期以来,视频背景作为视觉信息中的“隐形角色”,在视频处理任务中常被边缘化. 但在视频质量评估领域,背景信息却成为评估整体质量不可或缺的关键. 文献[7]中专门针对从失真角度对图像质量评价研究进行了综述,包含前景失真和背景失真;文献[8]中通过提取背景活动、背景亮度等信息,结合边缘信息一起,作为图像质量评价的关键因素;文献[9]中则考虑不同亮度背景的影响,即“亮度掩盖效应”,来研究基于亮度掩盖的图像质量评价模型.

为验证预训练网络在特征提取中可能忽视背景失真的假设,我们设计了一系列实验,采用故障监控视频帧与噪声视频帧为样本(如图 2 所示),结合 Grad_CAM[10]技术可视化特征热力图. 实验结果显示,网络在处理条纹失真监控视频时,未能聚焦于条纹,反被车辆吸引;在处理含有椒盐噪声的道路视频时,尽管捕捉到部分背景失真,但仍遗漏了大量背景细节,这有力支持了我们的假设.



(a) 失真视频帧　　(b) 提取的特征热力图

**图 2　失真视频帧及其热力图**

**Fig. 2　Distorted video frames and their heat maps**

针对上述问题,本文提出了一种考虑背景失真的无参考 VQA 方法,该方法旨在全面考虑目标内容与背景环境的综合影响,以提升视频质量评估的准确性. 具体创新点包括:

(1)构建特征丰富模块:在预训练好的 ResNet50 模型基础上,构建特征丰富模块,深入挖掘底层图像信息,以在高层语义特征的基础上进一步丰富特征,实现全面高效提取.

(2)引入通道挖掘机制:对增强后的特征精细筛选,保留浅层背景边缘细节,捕捉深层背景语义信息,实现特征信息全面优化.

(3)设计时序模块:基于 RNN 设计时序模块,兼顾长期与短期时序相关性,为空间特征提供精细时序建模能力,确保视频质量评估全面细致.

## 1　相关工作

现阶段的无参考视频质量评价(no reference video quality assessment,NR-VQA)研究聚焦于压缩视频、自然视频及用户生成内容(user generated content,UGC),分为传统方法与深度学习方法两大阵营[11]. 传统方法依托高斯分布等数学模型构建评价数据集的整体分布,并借助支持向量机训练通用模型以快速评估视频质量,然而,数据多样性挑战了其泛化能力. 相反,深度学习方法凭借数据驱动的优势,在模型泛化性

上展现出卓越表现,核心模型包括二维 CNN,结合 RNN 及三维 CNN 等.

具体而言,文献[4]将二维 CNN 与长短期记忆网络(long short-term memory,LSTM)相结合,通过分类网络提取高层语义特征捕捉内容信息,LSTM 则负责跨帧的语义特征长期依赖建模,并模拟人眼视觉系统的延迟效应,在自然视频数据集上实现了良好的主观相关性. 受其启发,文献[12]强调语义内容与人类感知的一致性,利用双 CNN 分别编码视频质量属性与语义内容,辅以 RNN 进行时序建模,进一步优化了质量评估. 文献[13]则聚焦于提升预测速度,通过创新的采样方法优化了[12]中的算法. 文献[14]则另辟蹊径,将动态变化与内容移动的混合失真纳入考量,融合增强动态信息感知于静态对象感知中,旨在提升泛化性与主观性.

此外,二维 CNN 作为骨干网络在视频质量评估中占据主流地位,但伴随的是高层语义特征中背景失真信息的损失,这在特定场景下限制了评估的准确性. 为此,文献[15]、[16]、[17]通过结合其他网络模块探索了对普通视频、声呐图像及低光照图像增强结果的质量评估;而文献[18]、[19]则致力于三维 CNN 的改进,通过多任务训练等策略,使其更适用于视频质量评价任务.

综上所述,尽管二维 CNN 是当前视频质量评价的主要架构,但如何有效保留并利用背景失真信息,仍是提升评估精度与广泛适用性的关键研究方向.

## 2　方法

本文聚焦于背景失真信息缺失的核心问题,通过以下策略进行系统性优化:首先,鉴于预训练残差网络各层特征信息的差异性,特别是浅层特征富含图像细节而语义信息较弱的特点[20],提出跨层特征补充机制,以弥补不同层级间的信息空白. 其次,采用通道挖掘技术,对特征通道进行精细化筛选,旨在低维特征中精准提取与背景失真紧密相关的边缘细节,同时在高维特征中筛选出关键语义信息,从而实现特征信息的优化整合. 最后,引入时序模块,对整合后的特征进行时间维度的深度处理,以捕捉视频序列中的动态变化信息.

所设计的模块结构如图 3 所示,该架构由三大核心组件构成:特征丰富模块(feature enrichment module,FEM),负责跨层特征融合以增强信息完整性;通道挖掘模块(channel mining module,CMM),实现特征通道的智能选择与优化;以及时序建模模块(time series modeling module,TSMM),负责在时间维度上解析视频特征,确保评估结果既准确又全面. 下面将分别对这三大模块进行详细的论述.

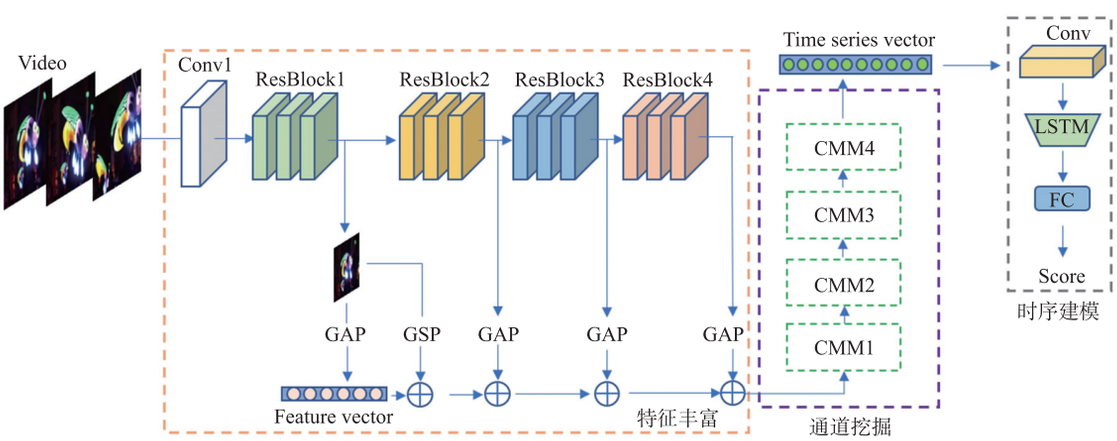



**图 3　整体网络结构图**

**Fig. 3　The overall network structure diagram**

### 2.1　特征丰富模块

该模块采用预训练的 ResNet50 作为特征提取的主干网络,从 ResNet50 中选取 ResBlock1 至 ResBlock4 这 4 个关键层的输出特征,对选取的特征进行自适应融合,旨在提高特征的表达能力和效率,同时通过融合后的特征,构建出更为高效的空间特征.

鉴于各层特征图尺寸不同,直接融合面临挑战. 因此,我们将全局标准池化(global standard pooling,GSP)[4]与全局平均池化[21](global average pooling,GAP)相结合,在此基础上对特征图进行降维和融合,

具体处理方法如下:

(1)针对 ResBlock1 的特征图,并行应用 GAP 和 GSP 技术,将特征图压缩至 1×1×$C$ 维度后进行拼接,形成初始特征向量 $F^{gsp}_{block\ 1}$.

(2)针对 ResBlock2、ResBlock3 和 ResBlock4 的特征图,依次采用 GAP 处理当前层特征. 将当前层处理得到的特征向量 $F^{gap}_{block\ i}$ 与前一层拼接得到的特性向量进行串联,逐步构建出特征金字塔.

值得注意的是,我们摒弃了 Conv1 层的输出特征,原因在于其较低的抽象层次可能引入噪声,干扰高层特征的语义纯度. 最终,特征丰富模块的输出构成经过精心配比:ResBlock4 在整体贡献中占据最大比例,为 1/2,ResBlock3 次之,占比 1/4,ResBlock2 占比 1/8,ResBlock1 通过两个不同视角的特征,贡献剩余的 1/8. 整个融合后的特征通道维度高达 4096,确保了信息的丰富性与表达的深度,具体实现细节如图 4 所示.




**图 4　特征丰富模块输出特征**

**Fig. 4　Output feature of FEM**

## 2.2　通道挖掘模块

特征丰富模块在增强信息丰富性的同时,也带来了冗余信息的挑战. 为此,我们设计了通道挖掘模块,该模块利用门控注意力机制智能筛选通道信息,确保输出特征在通道维度上既精确又具备更强的适应性.

针对高通道数特征处理中的挑战,传统全连接层往往导致参数数量急剧增加,从而导致模型复杂度及过拟合风险的增加. 尤其在本场景下,特征图尺寸为 1×1,已不具备空间上的实际意义. 因此,我们提出以下优化策略:

(1)特征维度转换:通过高效的张量操作,直接去除特征的空间维度,将其转换为仅包含时间维和通道维的二维数据. 这一步骤简化了特征结构,使其更加适应后续处理.

(2)深度挖掘与降维:随后,应用一维卷积层对转换后的二维数据进行深度挖掘,有效提取通道维度上的关键信息,并实现降维. 这一方法不仅保持了特征的有效性,还显著降低了计算复杂度,提升了模型的整体性能.

在具体实现时,通道挖掘模块由串联的 4 个子模块组成,每个子模块针对不同通道数进行精细挖掘. 如图 5 所示,每个子模块包含 4 个二维卷积层,前三层用于生成不同通道数的特征映射. 随后,利用基于 sigmoid 激活函数的门控注意力机制,结合背景边缘与语义信息,对这些特征进行深度挖掘. 挖掘后的特征通过矩阵点乘方式与原特征融合,增强注意力效果. 最终,将各子模块挖掘到的特征拼接,形成新的特征输入下一轮挖掘. 整个模块的输出是一个融合了通道与时间信息的二维向量,有效去除了冗余,保留了关键信息.

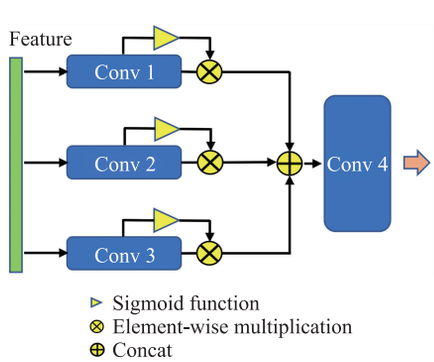



**图 5　通道挖掘子模块结构图**

**Fig. 5　Structure diagram of channel mining sub-module**

表 1 详细列出了每个子模块中卷积层的通道参数配置,明确展示了不同层的输入与输出通道数. 值得注意的是,所有卷积层均采用一维卷积实现,固定卷积核与步长均为 1,确保了操作的简洁与高效. 其中,$Sub_i$ 代表第 $i$ 个通道挖掘子模块,这一设计体现了模块化的灵活性与可扩展性.

**表 1　通道挖掘模块卷积层通道参数配置**

**Table 1　Configuration of channel parameters in convolution layer of CMM**

| | Params | $Sub_1$ | $Sub_2$ | $Sub_3$ | $Sub_4$ | | Params | $Sub_1$ | $Sub_2$ | $Sub_3$ | $Sub_4$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Conv1 | In channel | 4 096 | 1 024 | 256 | 64 | Conv2 | In channel | 4 096 | 1 024 | 256 | 64 |
| | Out channel | 1 024 | 256 | 64 | 16 | | Out channel | 512 | 128 | 32 | 8 |
| Conv3 | In channel | 4 096 | 1 024 | 256 | 64 | Conv4 | In channel | 1 792 | 448 | 112 | 28 |
| | Out channel | 256 | 64 | 16 | 4 | | Out channel | 1 024 | 256 | 64 | 16 |

### 2.3　时序建模模块

经过通道挖掘模块的处理,我们获得了最终的空间特征表示. 由于视频中的时间维度和空间信息同样重要,因此,我们设计了一个时序建模模块. 鉴于视频观看时,前一两秒的内容往往对当前帧的感知有显著影响,展现出长期的依赖性,因此,在利用 RNN 进行时间维度建模的基础上[4,6],选择 LSTM 来捕捉视频中的长期依赖效应. 同时,为捕捉当前帧与邻近帧间潜在的动态信息,我们还引入了一维卷积进行时序建模.

如图 6 所示,该时序模块由 4 个关键层组成:首先,采用卷积核大小为 5 且步长为 2 的一维卷积层,旨在高效捕获相邻五帧之间的信息联系. 考虑到相邻帧间变化细微,步长设为 2 不仅可视为对相似帧的合并处理,还能有效将特征序列的时间维度减半,从而优化内存使用并减少计算负荷. 随后,是 LSTM 层,专门用于对经过初步处理的特征进行长期依赖关系的深度建模. 最后,通过全连接(fully connected,FC)层进行特征映射,为每帧生成预估的质量分数. 这些分数随后被送入池化层[4],得出最终的视频质量评分.

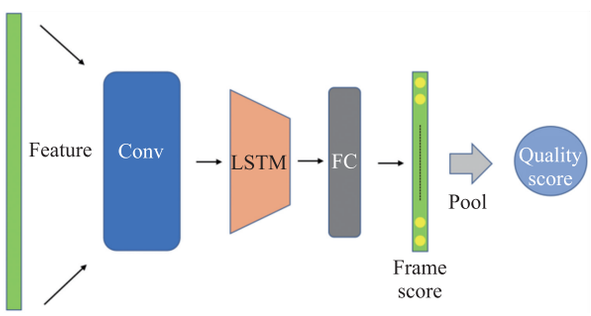



**图 6　时序建模模块结构图**

**Fig. 6　The structure diagram of TSMM**

## 3　实验

### 3.1　实验设置

#### 3.1.1　实验参数

实验依托深度学习框架 Pytorch 实现,硬件配置如下:

(1)Intel i7-8700K 处理器;

(2)NVIDIA GTX 2080 Ti 显卡(11GB 显存);

(3)32GB 内存.

我们参照文献[4]的方法设置时序建模模块中池化操作的超参数,采用随机梯度下降法,初始学习率设置为 0.0001,并进行了总共 2000 次的迭代训练,每次迭代处理的批次大小设置为 16.

#### 3.1.2　实验数据集

实验所使用的数据集为 KonViD-1k[22]、LIVE-Qualcomm[23]和 CVD2014[24]. KonViD-1k 由康斯坦茨大学研究人员提出,从公共视频中均等地收集了总共 1200 个视频,包含了各种自然失真现象;CVD2014 则是包含了一些在视频获取过程中产生的失真;LIVE-Qualcomm 则是由美国德克萨斯大学奥斯汀分校影像与视频工程实验室(Laboratory for Image & Video Engineering)建立,包含了 208 个失真视频. 数据集的详细信息如表 2 所示.

**表 2　VQA 数据集详细信息**

**Table 2　The details of VQA datasets**

| 数据集 | 来源 | 失真视频 | 场景内容数 | 采集设备数 | 失真数 | 分数 | 视频长度 | 分辨率 | 帧率(FPS) |
|---|---|---|---|---|---|---|---|---|---|
| LIVE Qual | 实验室 | 208 | 54 | 8 | 6 | 0-100 | 15s | 1920x1080 | / |
| CVD2014 | 实验室 | 234 | 5 | 78 | / | 0-100 | 10 s-25 s | 1 280×720 | 11-31 |
| KoNViD-1k | 众包 | 1 200 | 1 200 | >164 | / | 0-5 | 8 s | 960x540 | 24/30 |

鉴于自然视频数据集仅适合使用 NR-VQA 方法,我们选择了 VSFA[4]、VBLINDS[25]、VIDEO[26]、BRISQUE[27]、CORNIA[28]、QSA-VQM[13]、MDFFN[14]和 NIQE[29]这 7 个无参考评价方法,以进行对比分析. 其中,VSFA 为首个采用 CNN 模型对自然视频质量进行评估的方法.

#### 3.1.3　实验指标

本论文所采用的实验指标为:斯皮尔曼秩相关系数(spearman rank order correlation coefficient,SROCC)、肯德尔秩相关系数(Kendall rank order correlation coefficient,KROCC)、皮尔逊线性相关系数(Pearson linear correlation coefficient,PLCC)和均方根误差(root mean squared error,RMSE).

### 3.2 实验性能

在公开数据集 KoNViD-1k、LIVE-Qualcomm 和 CVD2014 上开展实验,采用 3∶1∶1 的配置,且训练、验证和测试集三部分无重叠. 随机划分可能导致质量好的视频集中于训练集或测试集,影响实验效果. 因此,本文采用随机划分的方法,取 10 次的平均值为最后结果,以减少偏差.

从表 3 和表 4 中可以看出,与基于传统方法的 NR-VQA 模型相比,本文方法在 3 个数据集上的性能均有提升. 而与近几年基于深度学习的 NR-VQA 模型 VSFA、MDFFN 和 QSA-VQM 相比,本文方法的整体效果最好.

**表 3 KonViD-1k 和 CVD2014 数据集上不同算法的性能**

**Table 3 Performance of different algorithms on KoNViD-1k and CVD2014 datasets**

| 算法 | KoNViD-1k | | | | CVD2014 | | | |
|---|---|---|---|---|---|---|---|---|
| | SROCC ↑ | KROCC ↑ | PLCC ↑ | RMSE ↓ | SROCC ↑ | KROCC ↑ | PLCC ↑ | RMSE ↓ |
| BRISQUE[14] | 0.654(±0.042) | 0.473(±0.034) | 0.626(±0.041) | 0.507(±0.031) | 0.709(±0.067) | 0.518(±0.060) | 0.715(±0.048) | 15.197(±1.325) |
| NIQE[29] | 0.544(±0.040) | 0.379(±0.029) | 0.546(±0.038) | 0.536(±0.010) | 0.489(±0.091) | 0.358(±0.064) | 0.593(±0.065) | 17.168(±1.138) |
| CORNIA[15] | 0.610(±0.034) | 0.436(±0.029) | 0.608(±0.032) | 0.509(±0.014) | 0.614(±0.075) | 0.441(±0.058) | 0.618(±0.079) | 16.871(±1.200) |
| VIDEO[14] | 0.298(±0.052) | 0.207(±0.035) | 0.303(±0.049) | 0.610(±0.012) | 0.023(±0.122.) | 0.021(±0.081) | −0.025(±0.144) | 21.822(±1.152) |
| VBLINDS[12] | 0.695(±0.024) | 0.509(±0.020) | 0.658(±0.025) | 0.483(±0.011) | 0.746(±0.056) | 0.562(±0.057) | 0.753(±0.053) | 14.292(±1.413) |
| MDFFN[18] | 0.789 | / | 0.788 | / | 0.847 | / | 0.849 | / |
| QSA-VQM[17] | 0.810 | / | 0.810 | 0.391 | 0.861 | / | 0.870 | 11.030 |
| VSFA[4] | 0.773(±0.019) | 0.578(±0.023) | 0.775(±0.020) | 0.421(±0.021) | 0.870(±0.037) | 0.695(±0.046) | 0.868(±0.031) | 10.857(±1.352) |
| Ours | 0.804(±0.030) | 0.612(±0.029) | 0.812(±0.033) | 0.451(±0.030) | 0.880(±0.024) | 0.701(±0.041) | 0.872(±0.027) | 11.543(±1.123) |

注:↑代表值越大主观相关性越高,↓代表值越小误差越小,效果越好;加粗字体表示最佳效果,加下划线表示次佳效果.

**表 4 LIVE-Qualcomm 数据集与整体数据集上不同算法的性能**

**Table 4 Overall performance and performance of different algorithms on LIVE-Qualcomm dataset**

| 算法 | LIVE-Qualcomm | | | | Overall Performance | | | |
|---|---|---|---|---|---|---|---|---|
| | SROCC ↑ | KROCC ↑ | PLCC ↑ | RMSE ↓ | SROCC ↑ | KROCC ↑ | PLCC ↑ | RMSE ↓ |
| BRISQUE[14] | 0.504(±0.147) | 0.365(±0.111) | 0.516(±0.127) | 10.731(±1.335) | 0.643(±0.059) | 0.465(±0.047) | 0.625(±0.053) | 3.895(±0.380) |
| NIQE[29] | 0.463(±0.105) | 0.328(±0.088) | 0.464(±0.136) | 10.858(±1.013) | 0.526(±0.105) | 0.369(±0.041) | 0.542(±0.054) | 4.214(±0.323) |
| CORNIA[15] | 0.460(±0.130) | 0.324(±0.104) | 0.494(±0.133) | 10.759(±0.939) | 0.591(±0.052) | 0.423(±0.043) | 0.595(±0.051) | 4.139(±0.300) |
| VIDEO[13] | 0.127(±0.137) | 0.082(±0.099) | −0.001(±0.106) | 12.308(±0.881) | 0.237(±0.073) | 0.164(±0.050) | 0.218(±0.070) | 5.115(±0.285) |
| VBLINDS[12] | 0.566(±0.078) | 0.405(±0.074) | 0.568(±0.089) | 10.760(±1.231) | 0.686(±0.035) | 0.503(±0.032) | 0.660(±0.037) | 3.753(±0.365) |
| MDFFN[18] | 0.807 | | 0.823 | | 0.814 | | 0.820 | |
| QSA-VQM[17] | 0.740 | | 0.770 | 7.930 | 0.803 | | 0.816 | 6.450 |
| VSFA[4] | 0.745(±0.037) | 0.559(±0.046) | 0.748(±0.031) | 8.542(±1.352) | 0.771(±0.028) | 0.582(±0.029) | 0.762(±0.031) | 3.704(±0.448) |
| Ours | 0.763(±0.033) | 0.587(±0.042) | 0.752(±0.037) | 8.482(±1.245) | 0.815(±0.029) | 0.633(±0.056) | 0.811(±0.032) | 8.482(±0.799) |

在评估算法性能时,我们观察到 CVD2014 数据集上算法表现普遍出色,这得益于其数据量相对适中,仅为 KoNViD-1k 数据集的五分之一. 然而,当面对 LIVE-Qualcomm 数据集时,由于该数据集复杂性显著提升,算法性能出现了一定程度的下滑,表现接近于 KoNViD-1k 的水平. 深入分析发现,CVD2014 数据集的 RMSE 值是 KoNViD-1k 的两倍,这主要是由于其数据量较小,容易导致模型过拟合,进而增大了预测误差. 相比之下,LIVE-Qualcomm 数据集的内容更为复杂,使得模型难以达到最优拟合状态,预测结果略显不准确.

尽管如此,在两个数据集上,算法的相关性仍然保持在高水平,视频质量等级的预测也相当准确. 在与 VSFA 算法的比较中,我们发现本文提出的方法在 3 个数据集上的总体 RMSE 值稍高. 这主要是因为通道挖掘模块虽然进行了通道选择,但卷积层的叠加增加了模型的参数量,使得模型在训练过程中难以完全拟合,从而导致预测值与 MOS 值之间的误差略大. 不过,值得注意的是,其他主观相关性指标仍然显示出我们的模型具有良好的主观性能.

最后,我们使用本文提出的模型对 3 个数据集中的测试视频进行了预测,结果如图 7 所示. 与文献[4]中的方法相比,我们的模型预测的质量分数更加接近 MOS 值,这进一步验证了模型的有效性.




**图 7　模型预测视频的主观质量分数与真实质量分数**

**Fig. 7　The predicted score and MOS of our model and VSFA**

为验证论文方法在评价交通监控视频/图像质量时的有效性,对一些交通监控图像进行了评价,如图 8 所示. 从图 8 中可以看出,图(c)中背景失真最少,因而其分数最高,而图(b)、(d)、(e)、(f)中背景灯光导致的失真影响了整体质量的评价,使其分数较低,这 4 个里面,(b)、(e)的背景失真稍微少一些,因而其分数相对高一点. 而图(a)中,图像整体处于模糊失真状态,因而其分数最低. 从图 8 也可以看出,评分的情况与主观的观感是保持一致的.

### 3.3　消融实验

消融实验在 KoNViD-1k 数据集上开展,实验效果如表 5 所示. 基础模型为文献[4]中所述的 VSFA 模型,我们使用本文提出的 3 个模块分别替换基础模型中对应的网络部分来进行实验. 对表 5 数据进行分析可知,实验效果最佳的是替换特征丰富模块,SROCC 和 PLCC 值均接近 0.8,显示出较高的主观相关性. 同时,较小的均方根误差值也表明预测的主观分值与 MOS 值相当接近. 而仅替换通道挖掘模块的实验验证了先前的假设,即通道挖掘模块能提供更好的主观相关性,然而,在预测准确度方面,该模块的表现略显不足. SROCC、PLCC 和 RMSE 都大于单独替换特征丰富模块,最后在只替换时序模块时,总体性能较不使用时序模块时均有所提升,误差也较小. 使用时序模块与特征丰富模块的组合替换能够减少误差值并略微提升相关性;使用时序与通道挖掘模块替换虽然相关性相比于只使用通道挖掘模块替换时略微下降,但误差得到了明显改善.




**图 8　针对交通监控视频的主观质量评价分数**

**Fig. 8　The predicted scores of our method on traffic monitoring videos**

**表 5　KoNViD-1k 上不同模块实验性能**

**Table 5　Performance of different modules on KoNViD-1k dataset**

| FEM | CAM | TSMM | SROCC ↑ | KROCC ↑ | PLCC ↑ | RMSE ↓ |
|---|---|---|---|---|---|---|
| | | | 0.773 | 0.578 | 0.775 | 0.421 |
| √ | | | 0.798 | 0.601 | 0.798 | 0.401 |
| | √ | | 0.801 | 0.610 | 0.803 | 0.616 |
| | | √ | 0.788 | 0.591 | 0.791 | 0.409 |
| √ | | √ | 0.795 | 0.604 | 0.808 | 0.381 |
| √ | √ | | 0.799 | 0.607 | 0.795 | 0.424 |
| | √ | √ | 0.794 | 0.603 | 0.806 | 0.474 |
| √ | √ | √ | 0.804 | 0.612 | 0.812 | 0.451 |

从整体来看,同时使用 3 个模块替换网络对应部分,能极大改善单独替换时的性能不足. 在 3 个模块共同作用下,KoNViD-1k 数据集上的 SROCC 和 PLCC 均提升至 0.8 以上,KROCC 达到 0.6 以上. 尽管均方根误差相比 VSFA 有所升高,但总体上仍保持在一个较低的误差水平.

为了验证特征丰富模块中 ResBlock1、2、3 和 4 的特征组合是否最有效,我们对此进行了验证,模型使用了通道挖掘与时序模块与不同 ResBlock 组合的特征进行实验,实验最终结果如表 6 所示,ResBlock1 和 2 组合的特征与 ResBlock4 单独的语义特征相比,其在 SROCC 等主观相关性指标上都有较大的下降,这说明了高层的语义特征在视频质量评价任务中仍然十分重要. 而将 ResBlock1 和 3 分别与 4 进行组合后相比于 4 整体仍有下降,但在加入 ResBlock2 输出特征后整体性能有了小幅提升,因此我们认为 ResBlock2 对于背景失真特征的捕捉更加完善. 最后再组合 ResBlock1 性能仍有略微提升. 鉴于 ResBlock1 特征通道数较小,仅为 32,因此其不会增加太大的计算量,却能够带来性能提升,最终本文使用了 ResBlock1、2、3 和

4 组合的特征作为特征丰富模块的输出.

表 6 不同 ResBlock 输出特征融合后的实验性能

Table 6 Performance of fusing features from different ResBlocks

| Module | SROCC | KROCC | PLCC | RMSE | Module | SROCC | KROCC | PLCC | RMSE |
|---|---|---|---|---|---|---|---|---|---|
| ResBlock 4 | 0.794 | 0.603 | 0.806 | 0.474 | ResBlock 1,2 and 4 | 0.798 | 0.605 | 0.809 | 0.482 |
| ResBlock 1 and 2 | 0.771 | 0.572 | 0.775 | 0.503 | ResBlock 2,3 and 4 | 0.802 | 0.608 | 0.809 | 0.454 |
| ResBlock 1 and 4 | 0.787 | 0.595 | 0.798 | 0.473 | ResBlock 1,2,3 and 4 | 0.804 | 0.612 | 0.812 | 0.451 |
| ResBlock 3 and 4 | 0.792 | 0.599 | 0.803 | 0.465 | | | | | |

此外,为验证预训练参数鲁棒性问题[6],在 KoNViD-1k 数据集上开展实验,构建无 ResNet50 线性回归任务预训练任务. 如表 7 所示,实验结果表明,KoNViD-1k 数据集预训练能提升在该数据集上的性能,但在其他数据集上的效果却明显下降. 而 ImageNet 预训练参数在 3 个数据集上性能均佳. 上述结果说明,预训练效果与数据集关联较大;此外,ImageNet 预训练参数鲁棒性更好,更加有效,故被本文采用.

表 7 不同数据集上不同预训练参数实验效果

Table 7 Performance of different pretraining parameters on different datasets

| Module | Network parameter | SROCC | KROCC | PLCC | RMSE |
|---|---|---|---|---|---|
| KoNViD-1k | KoNViD Pretrained | 0.813 | 0.631 | 0.823 | 0.391 |
| | ImageNet Pretrained | 0.804 | 0.612 | 0.812 | 0.451 |
| CVD2014 | KoNViD Pretrained | 0.851 | 0.653 | 0.860 | 15.423 |
| | ImageNet Pretrained | 0.880 | 0.701 | 0.879 | 11.543 |
| LIVE-Qualcomm | KoNViD Pretrained | 0.751 | 0.563 | 0.726 | 10.122 |
| | ImageNet Pretrained | 0.763 | 0.587 | 0.752 | 8.482 |

# 4 结论

在进行自然拍摄视频质量评价时,综合考虑目标内容与背景环境的共同影响,提出了一种 NR-VQA 方法,该方法使用特征丰富模块来捕获更多前景内容和背景失真信息,使用通道挖掘机制来提取更重要通道中的信息,并丢弃冗余的通道信息,使用时序建模模块有效地捕捉时间维关系,同时解决高维度特征较难训练的问题.

实验结果表明,与主流的 NR-VQA 方法相比,本文所提评价方法提高了 VQA 的整体性能. 未来工作将进一步对特征的高效提取进行研究,并尝试改善损失函数以降低模型的评价误差,实现更好的预测效果.

## [参考文献]


[1] KOU T,LIU X,SUN W,et al. Stablevqa:a deep no-reference quality assessment model for video stability[C]//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa,Canada,2023:1066-1076.

[2] SUN W,MIN X,LU W,et al. A deep learning based no-reference quality assessment model for ugc videos[C]//Proceedings of the 30th ACM International Conference on Multimedia. Lisbon,Portugal,2022:856-865.

[3] ZHANG Z,SUN W,ZHU Y,et al. Evaluating point cloud from moving camera videos:A no-reference metric[J]. IEEE transactions on multimedia,2023:1-13.

[4] LI D,JIANG T,JIANG M. Quality assessment of in-the-wild videos[C]//Proceedings of the 27th ACM International Conference on Multimedia. Nice,France:Association for Computing Machinery,2019:2351-2359.

[5] LI D,JIANG T,JIANG M. Unified quality assessment of in-the-wild videos with mixed datasets training[J]. International journal of computer vision,2021,129(4):1238-1257.

[6] VARGA D,SZIRáNYI T. No-reference video quality assessment via pretrained CNN and LSTM networks[J]. Signal,image and video processing,2019,13(8):1569-1576.

[7] 鄢杰斌,方玉明,刘学林. 图像质量评价研究综述——从失真的角度[J]. 中国图像图形学报,2022,27(5):1430-1466.

[8] BABU R V,SURESH S,PERKIS A. No-reference JPEG-image quality assessment using GAP-RBF[J]. Signal processing,2007,87(6):1493-1503.
[9] 王怡影,李琛璐,柴豆豆 等. 基于亮度掩盖的图像质量评价模型[J]. 新乡学院学报,2021,38(3):38-41.
[10] RAMPRASAATH R S,MICHAEL C,ABHISHEK D,et al. Grad-CAM:Visual explanations from deep networks via gradient-based localization[C]//Proceedings of 2017 IEEE International Conference on Computer Vision(ICCV). Venice,Italy:IEEE,2017:618-626.
[11] 程茹秋,余烨,石岱宗 等.图像与视频质量评价综述[J]. 中国图像图形学报,2022,27(5):1410-1429.
[12] AGARLA M,CELONA L,SCHETTINI R. No-reference quality assessment of in-capture distorted videos[J]. Journal of imaging,2020,6(8):74.
[13] AGARLA M,CELONA L,SCHETTINI R. An efficient method for no-reference video quality assessment[J]. Journal of imaging,2021,7(3):55.
[14] JIANG J,WANG X,LI B,et al. Multi-dimensional feature fusion network for no-reference quality assessment of in-the-wild videos[J]. Sensors,2021,21(16):5322.
[15] 姚军财,申静,黄陈蓉. 基于多层 BP 神经网络的无参考视频质量客观评价[J]. 自动化学报,2022,48(2):594-607.
[16] 李钟华,陈禅,王雪津,等. 基于跨网络特征补偿的声呐图像质量评价方法[J]. 北京航空航天大学学报,网络首发 2024-10-16. DOI:10.13700/j.bh.1001-5965.2024.0440.
[17] YU Y,CHEN F,JU J,et al. LMT-GP:Combined latent mean-teacher and gaussian process for semi-supervised low-light image enhancement[C]//Proceeding of the 2024 European Conference on Computer Vision(ECCV). Milan,Italy,2024:261-279.
[18] LIU W T,DUANMU Z F,WANG Z. End-to-end blind quality assessment of compressed videos using deep neural networks [C]//Proceedings of the 26th ACMet international conference on Multimedia. Seoul South Korea:ACM,2018:546-554.
[19] 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[J]. 软件学报,2017,27(S2):103-112.
[20] 余烨,傅云翔,杨昌东,等. 基于 FR-ResNet 的车辆型号精细识别研究[J]. 自动化学报,2021,47(5):1125-1136.
[21] YU Y,LIU H,FU Y,et al. Embedding pose information for multiview vehicle model recognition[J]. IEEE transactions on circuits and systems for video technology(TCSVT),2022,32(8):5467-5480.
[22] HOSU V,HAHN F,JENADELEH M,et al. The konstanz natural video database(KoNViD-1k)[C]//Proceedings of 2017 Ninth International Conference on Quality of Multimedia Experience(QoMEX). Erfurt:IEEE,2017:1-6.
[23] GHADIYARAM D,PAN J,BOVIK A C,et al. In-capture mobile video distortions:a study of subjective behavior and objective algorithms[J]. IEEE transactions on circuits and systems for video technology,2018,28(9):2061-2077.
[24] NUUTINEN M,VIRTANEN T,VAAHTERANOKSA M,et al. CVD2014—A database for evaluating no-reference video quality assessment algorithms[J]. IEEE transactions on image processing,2016,25(7):3073-3086.
[25] SAAD M A,BOVIK A C,CHARRIER C. Blind prediction of natural video quality[J]. IEEE transactions on image processing,2014,23(3):1352-1365.
[26] MITTAL A,SAAD M A,BOVIK A C. 2016. A completely blind video integrity oracle[J]. IEEE transactions on image processing,2016,25(1):289-300.
[27] MITTAL A,MOORTHY A K,BOVIK A C. No-reference image quality assessment in the spatial domain[J]. IEEE transactions on image processing,2012,21(12):4695-4708.
[28] XU J T,YE P,LIU Y,DOERMANN D. No-reference video quality assessment via feature learning[C]//Proceedings of 2014 IEEE International Conference on Image Processing(ICIP). Paris,France:IEEE,2014:491-495.
[29] MITTAL A,SOUNDARARAJAN R,BOVIK A C. Making a “completely blind” image quality analyzer[J]. IEEE signal processing letters,2013,20(3):209-212.

[责任编辑:陆炳新]