

卷积与自注意力结合和时空融合的无参考视频质量评价

牛玉贞^{1,2,3)}, 王镇隆¹⁾, 张博林¹⁾, 陈羽中^{1,2,3)}

¹⁾ (福州大学计算机与大数据学院 福州 350108)

²⁾ (福建省网络计算与智能信息处理重点实验室(福州大学) 福州 350108)

³⁾ (大数据智能教育部工程研究中心 福州 350108)

(nyzhen1982@fzu.edu.cn)

摘要: 随着互联网上的视频分享日益普遍, 对视频进行质量评估越来越重要。现有的视频质量评价研究中, 卷积被广泛应用于视频的空域和时域特征提取, 但由于其固有的局部感知特性, 缺乏对于视频全局信息的感知; 同时, 已有方法未对空域和时域特征进行充分融合。针对上述问题, 提出一种结合卷积与自注意力机制和时空融合的无参考视频质量评价方法。为了提取更有效的空域和时域特征, 首先分别设计空域特征提取模块和时域特征提取模块, 其中, 空域特征提取模块通过结合卷积和自注意力的特征感知增强模块, 对卷积提取的局部特征通过长距离依赖进行全局特征增强, 时域特征提取模块联合视频帧序列和视频帧差序列, 通过 Video Swin Transformer 和 3D 卷积学习视频的时域特征; 然后使用时空特征融合增强模块, 通过自注意力机制对时域和空域特征进行融合并增强; 最后对视频质量进行预测。在 KoNVID-1k 和 YouTube-UGC 数据集上的实验结果表明, 与文中对比方法中最优的视频质量评价方法相比, 所提方法具有更出色的表现, 并通过消融实验验证了每个模块的有效性。

关键词: 无参考视频质量评价; 3D 卷积神经网络; 自注意力机制

中图分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2023-00808

No-Reference Video Quality Assessment Based on Convolution and Self-Attention Combined Network and Spatial-Temporal Fusion

Niu Yuzhen^{1,2,3)}, Wang Zhenlong¹⁾, Zhang Bolin¹⁾, and Cheng Yuzhong^{1,2,3)}

¹⁾ (College of Computer and Data Science, Fuzhou University, Fuzhou 350108)

²⁾ (Fujian Key Laboratory of Network Computing and Intelligent Information Processing (Fuzhou University), Fuzhou 350108)

³⁾ (Engineering Research Center of Big Data Intelligence, Ministry of Education, Fuzhou 350108)

Abstract: Sharing videos on the Internet is becoming increasingly common, therefore, video quality assessment is becoming increasingly important. In existing video quality assessment research, convolution is widely used for feature extraction in the spatial and temporal domains of video, but due to its inherent local perceptual characteristics, the global information perception for video is neglected. At the same time, existing methods do not fully integrate spatial and temporal features. To solve the above problems, this paper proposes a no-reference video quality assessment method that combines convolution with a self-attention mechanism and spatiotemporal fusion. To extract more effective spatial and temporal features, the method first designs a spatial feature extraction module and a temporal feature extraction module, where the spatial feature extraction module enhances the global information of the local features extracted by convolution through long-range dependencies by the proposed feature perception and enhancement module that combines

收稿日期: 2023-12-18; 修回日期: 2024-07-15. 基金项目: 国家自然科学基金面上项目(61972097); 福建省自然科学基金(2023J01067). 牛玉贞(1982—), 女, 博士, 教授, 博士生导师, CCF 会员, 主要研究方向为计算机视觉、人工智能; 王镇隆(1998—), 男, 硕士研究生, 主要研究方向为深度学习、视频质量评估; 张博林(2000—), 男, 硕士研究生, 主要研究方向为深度学习、图像恢复; 陈羽中(1979—), 男, 博士, 教授, 博士生导师, 主要研究方向为计算智能、数据挖掘。

convolution and self-attention. The temporal feature extraction module jointly learns the temporal features of video using Video Swin Transformer and 3D convolution based on video frame sequences and video frame difference sequences. Then, the spatiotemporal feature fusion and enhancement module is used to fuse and enhance the temporal and spatial features through the self-attention mechanism, and finally predict the video quality. Experimental results on the KoNVid-1k and YouTube-UGC datasets show that the proposed method has better performance compared to existing video quality assessment methods, and ablation experiments also verify the effectiveness of each proposed module.

Key words: no-reference video quality assessment; 3D convolutional neural network; self-attention mechanism

近年来,随着科技的发展,通信网络的网速得到提升、智能手机不断推出、互联网视频平台兴起,人们能随时随地拍摄视频并在社交媒体上分享,导致视频的数量迅速增长。由于拍摄设备和拍摄经验与知识的差异,用户拍摄视频的视觉质量差异很大,因失焦、物体移动、相机抖动或曝光不足/过度等原因,拍摄出的视频呈现较低的视觉质量,给用户带来很差的视觉体验^[1]。因此,无论在拍摄过程中,还是在社交媒体的上传、审核与发布过程中,对拍摄的视频进行质量评价变得尤为重要。

根据对原始视频的参考程度,可以将视频质量评价任务分为无参考、半参考和全参考视频质量评价3类。由于实际应用场景中的视频一般没有参考视频,因此无参考视频质量评价的应用最广泛,也是目前视频质量评价模型的研究重点。

无参考视频质量评价方法可以分为传统方法和基于深度学习的方法^[2]。传统方法通过提取质量感知特征模拟空域和时域失真,常见的包括自然场景统计特征^[3-4]和运动特征^[5];但是传统方法泛化能力弱,只能针对特定失真类型的视频,不适用于情景更复杂的视频。基于深度学习的方法指基于深度神经网络的无参考视频质量评价模型,该方法自动地学习视频的视觉特征,且具有强大的特征表征能力和泛化能力,在性能上总体优于传统方法。

但是,现有的基于深度学习的方法仍然存在2个问题。(1) 无法对视频的空域和时域的局部和全局信息同时进行有效建模。已有的图像质量评价方法已经验证了全局信息的重要性^[6],而视频帧之间的长距离依赖关系能够反映用户观看具有不同视觉质量的帧序列时对整个视频的感知,因此空域和时域的全局信息对获得与用户感知一致的视频质量评价至关重要。已有的方法大多基于预训练的深层卷积神经网络。基于 ImageNet^[7]上预训

练的 ResNet50^[8]使得网络过多关注帧内的主体信息,无法获得影响最终视频质量的全局信息。(2) 无法充分地融合视频的时域和空域特征。已有方法中,研究者将时域特征和空域特征提取出来,然后通过简单的拼接或相加将其进行融合,再进行下一步的质量预测;但这种方法不能充分地利用时空特征之间的联系。为了更有效地利用所获得的时域特征和空域特征信息,需要设计一种更高效的时空特征融合方法。

为了解决上述2个问题,本文提出一种卷积与自注意力结合和时空融合的视频质量评价方法。(1) 分别设计空域特征提取模块和时域特征提取模块,以提取更有效的空域和时域特征。为了获取结合局部和全局信息的空域特征,在空域特征提取模块中包含一个结合卷积与自注意力的特征感知增强模块,用于进一步增强主干网络提取的空域特征。通过卷积的方式获取空域图像主体特征信息,联合自注意力机制获取空域背景即全局信息;再通过自适应权重将2种特征信息融合起来,使得空域特征不仅能够关注局部的空域信息,也能包含更多的全局信息用于后续的质量评估。时域特征提取模块联合整个视频的运动信息与视频帧间的残差信息,通过3D卷积和VST(video swin Transformer)^[9]学习视频的时域特征,其中,VST能够有效地提取视频帧之间的长距离依赖关系。(2) 为了更好地融合视频的时域特征和空域特征,本文提出一个基于自注意力机制的时空融合增强模块。通过自注意力的方式将时域特征与空域特征融合,再将融合特征与空域特征进行残差连接,使得融合信息能够进一步加强空域特征,获得更强的特征表征能力。

本文通过上述3个模块解决无参考视频质量评价存在的2个问题,为视频提供更加准确的质量分数。

1 相关工作

1.1 传统的无参考视频质量评价

早期的无参考视频质量评价方法中, 首先使用无参考图像质量评价方法^[10-11]计算每帧的质量分数, 然后计算整个视频的质量分数。但是, 由于视频的时域信息在视频质量评价的过程中起到很大作用, 导致直接采用上述过程的方法的性能有限。因此, 一些研究者使用视频的自然场景统计特征进行视频质量评价。Saad 等^[12]提出一种时空自然场景统计模型, 通过量化帧差和运动相干特征的自然场景统计特征实现视频质量评估; Mittal 等^[4]提出了一种名为 VIIDEO 的无参考视频质量评价模型, 利用视频的内在统计规律量化失真带来的干扰; Korhonen^[5]提出 TLVQM 模型, 更进一步地提取丰富的时空特征, 如运动、抖动、模糊、噪声、块状、颜色等几十种特征; Tu 等^[13]提出 VIDEVAL 模型, 其中结合了典型方法中的几十种有效特征进行训练。

视频内容对视频质量具有重要的影响, 虽然理解视频内容有利于视频质量评价, 但是以往的基于手工设计特征的方法很难理解语义信息。因此, 一些研究者将手工特征与预训练的卷积神经网络模型提取的语义级特征相结合, 以提高无参考视频质量评价模型的性能。Korhonen 等^[14]提出 C-TLVQM 模型, 其结合了 TLVQM^[5]中的手工特征和为图像质量评价预训练的卷积神经网络模型 ResNet50^[8]提取的空域特征; Tu 等^[15]提出了 RAPIQUE 模型, 通过结合质量感知的场景统计特征和语义感知的 ResNet50 提取的特征预测质量分数。C-TLVQM 和 RAPIQUE 模型均通过支持向量回归, 并将结合后的特征映射为质量分数。

1.2 深度学习的无参考视频质量评价

随着深度学习的发展以及大规模的无参考视频质量评价数据集^[16-17]的发布, 研究人员开始侧重对基于深度学习的无参考视频质量评价方法的研究。VSFA^[18]结合人类视觉系统的 2 个重要影响——内容依赖和时间记忆效应, 使用预训练的深层神经网络作为特征提取器, 再用门控循环单元进行时域信息建模, 最后使用时间滞后模型完成质量预测; Varga 等^[19]提出使用长短期记忆网络对时域信息建模的模型; Shahreza 等^[20]提出使用双向长短期记忆网络的模型。

近几年, 通过网络提取时域维度的信息并利用时域和空域特征获得视频的质量分数, 用于评

估视频质量已成为研究热点。RIRNet^[21]首先提取不同时间频率下的运动信息, 再将得到的运动信息通过分层失真描述量化时间运动效应, 从而对视频进行质量评估; Hou 等^[22]使用 3D 卷积神经网络进行视频质量评价; Ying 等^[23]和 Sun 等^[24]均通过从 2D 卷积神经网络中提取的空域特征和从 3D 卷积神经网络提取的时空特征进行联合质量评估; Xu 等^[25]首先从预训练的图像质量评价模型中提取视频帧的空域特征, 利用图卷积网络提取和增强这些特征, 然后从光流域中提取运动信息, 最后通过双向长短期记忆网络集成空域特征和运动信息; Li 等^[26]利用在多个数据库上预先构建的图像质量评估模型提取空域特征, 并利用动作识别模型提取时域特征, 然后利用门控循环单元网络对空域和时域特征进行建模; Zhang 等^[27]在时域特征和空域特征的基础上, 加入模糊、噪声、块效应、曝光和色彩 5 种手工设计的特征用于质量评价; Zhang 等^[28]提出 HVS-5M 模型, 进一步整合了边缘检测、内容感知、运动感知、时间滞后和视觉显著性 5 个部分对视频进行质量评价。

上述方法基于卷积神经网络、循环时间网络和 3D 卷积神经网络等, 比传统方法拥有更强的学习能力和适用性, 获得了更高的质量评价准确率; 但是, 这些方法存在无法对视频的局部和全局信息同时进行有效建模, 且无法充分地融合时空特征的问题。为此, 本文提出一种卷积与自注意力结合和时空融合的无参考视频质量评价方法, 取得了更好的效果。

2 本文方法

为了得到更加丰富、有效的视频空域和时域特征, 更好地融合时空特征用于准确评估视频质量, 本文提出一种卷积与自注意力结合和时空融合的无参考视频质量评价方法。该方法由空域特征提取模块、时域特征提取模块和时空特征融合增强模块 3 部分组成, 其整体结构如图 1 所示。空域特征提取模块包含一个预训练好的 ResNet50^[8]、特征感知增强模块和 2 个池化层。首先经过 ResNet50 提取第 5 个阶段输出的特征图, 通过特征感知增强模块从特征图获取更多视频的全局信息; 然后使用全局平均池化(global average pooling, GAP)和全局标准差池化(global standard deviation pooling, GSP)^[18]对特征进行特征整合, 得到最终的空域特征; 在时域特征提取模块中, 为了捕捉时域维度的

失真, 如闪烁、快门和焦点变化, 使用在 Kinetics 数据集^[29]上预训练的 VST 提取时域特征; 由于视频画面中的运动信息通常更容易吸引用户的注意力, 因此对视频进行帧差处理后送入一个 3D 卷积神经网络, 以获取视频运动特征, 再将 2 种特征进

行拼接组成时域特征图, 而后再使用 GAP 和 GSP 进行特征整合, 得到最终的时域特征. 最后将空域特征和时域特征送入时空融合增强模块, 通过自注意力机制将 2 种特征进行融合增强, 并利用质量回归模块得到最终的视频质量分数.

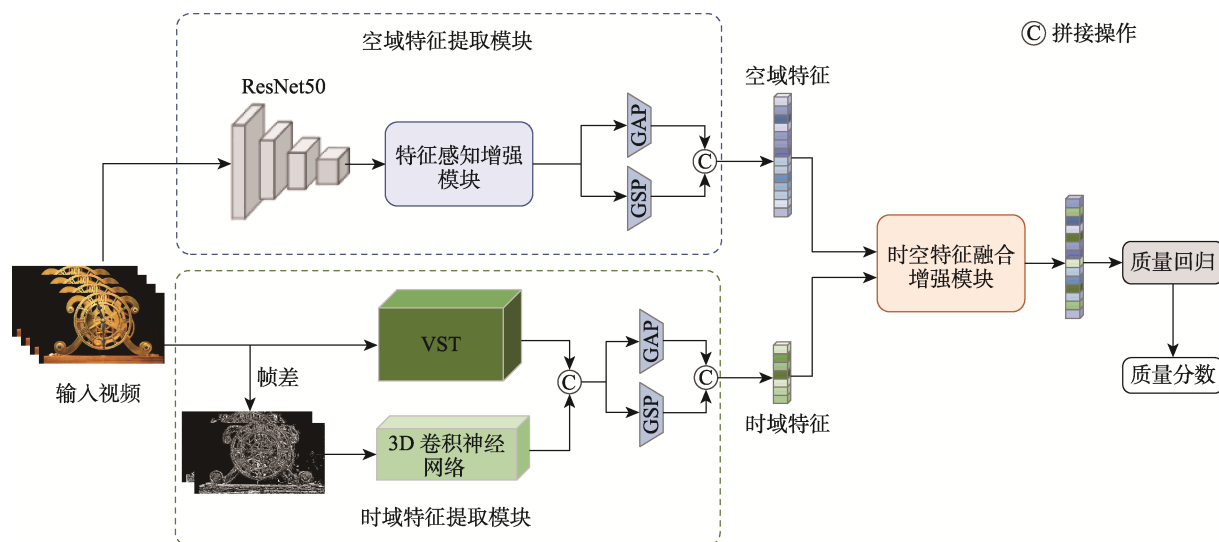


图 1 本文方法整体结构

2.1 特征感知增强模块

人类视觉系统中, 不同的视频内容会影响视频感知质量, 已有工作中大多使用预训练的深层卷积神经网络在空域上进行内容提取. 但是, 卷积的局部感知特性使得模型更关注视频中的内容主体信息, 难以提取全局信息; 而在视频质量评价任务中, 全局信息(包括背景信息、主体和背景之间的关系等)同样会对人的感官产生影响. 文献[30-33]验证了自注意力机制在捕获图像全局信息

方面的出色表现. 为了同时提取视频帧的局部和全局空域特征, 本文设计了基于卷积和自注意力的特征感知增强模块, 其结构如图 2 所示. 为了加快运算速度, 首先将 ResNet50 得到的深层特征图通过 3 个卷积层进行通道降维, 得到 3 个维度相同的特征图 $Q, K, V \in \mathbb{R}^{H \times W \times C}$; 然后通过自注意力和卷积分别对这 3 个特征图进行全局和局部特征提取.

在自注意力操作中, 先按照 $4 \times 4 \times C$ 的大小将 3 个 3D 的特征图 Q, K, V 分块, 并展开为 2D

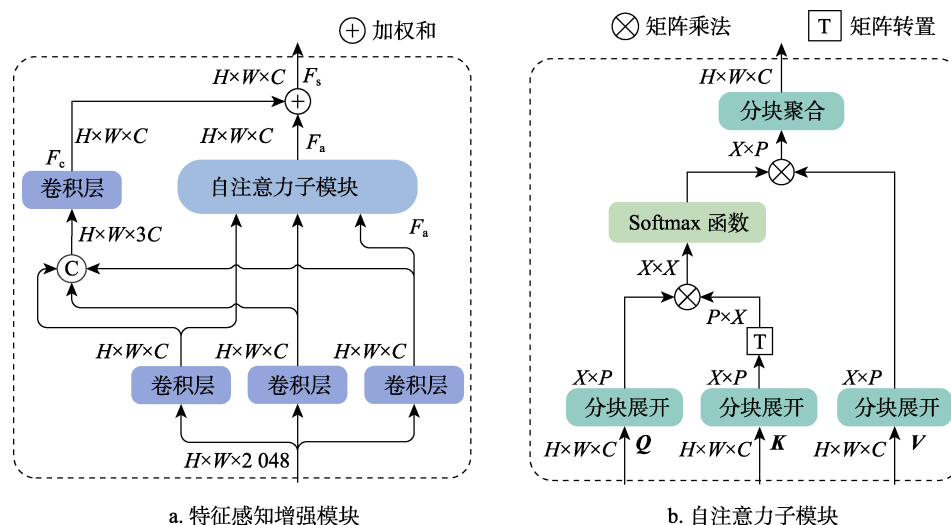
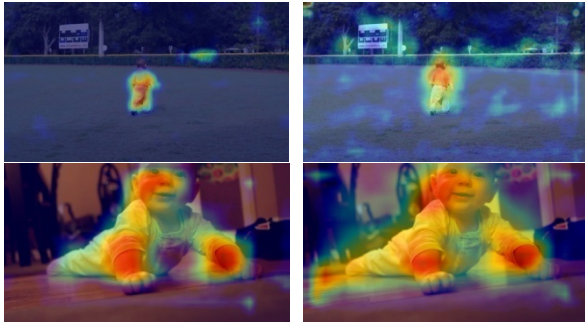


图 2 特征感知增强模块结构

特征 $\bar{Q}, \bar{K}, \bar{V} \in \mathbb{R}^{X \times P}$; 其中, $X = H \times W / (4 \times 4)$, $P = 4 \times 4 \times C$; 然后将这 3 个 2D 特征进行自注意力运算, 捕捉到空域中所需的全局特征信息; 最后将其变形成 $H \times W \times C$ 维度的视频空域全局特征. 卷积操作中, 先将这 3 个特征图 Q, K, V 进行拼接, 为了与全局特征保持相同的维度, 通过卷积进行降维, 得到 $H \times W \times C$ 维度的视频空域局部特征; 然后运用自适应加权的方法得到融合局部和全局的空域特征图; 最后经过 GAP 和 GSP 整合, 得到最终的空域特征 $f_s = \text{Cat}(\text{GAP}(F_s), \text{GSP}(F_s))$. 其中, $F_s = \alpha F_a + \beta F_c$ 表示通过特征感知增强模块得到的视频空域特征; $F_a = \text{Softmax}(\bar{Q} \circ \bar{K}^T) \circ \bar{V}$ 表示通过自注意力获得的视频空域全局特征; $F_c = \text{Conv}(\text{Cat}(Q, K, V))$ 表示通过卷积获得的视频空域局部特征; $\circ$ 表示矩阵乘法; $\text{Conv}(\cdot)$ 表示卷积操作; $\text{Cat}(\cdot)$ 表示拼接操作; α 和 β 表示可学习权重, 初始值都设为 0.5.

图 3 所示为特征感知增强模块和普通卷积模块得到的特征的热力图. 可以看出, 本文提出的特征感知增强模块可以增加模型对视频帧的全局注意力, 而不局限于视频的主体信息, 从而增加了特征的表征能力和模型的预测能力.



a. 卷积特征的热力图 b. 特征感知增强模块的热力图

图 3 普通卷积特征和特征感知增强模块得到的特征的热力图

2.2 时域特征提取模块

时域特征提取模块的目的是获取视频在时域维度的相关信息, 如视频抖动和运动模糊等. 传统的视频质量评价方法一般从帧差中获取视频的运动相关特征; 基于深度学习的方法则倾向于使用动作识别任务的预训练模型捕获时序失真信息, 如在 Kinetics 数据集上预训练的 3D ResNet-18^[34].

为了适应更复杂的视频运动情况, 本文使用基于深度学习的方法对视频时域信息进行建模,

考虑视频中的运动内容对于人类视觉系统的吸引力, 联合使用视频帧序列和视频帧差序列获得视频的时域信息.

针对模型大小以及效果, 使用在 Kinetics 数据集上预训练的 VST 基于视频帧序列提取视频的全局时域特征; 鉴于运动信息的重要性, 搭建了一个 3D 卷积神经网络, 基于视频帧差序列提取视频的局部时域特征, 该网络的结构如表 1 所示, 其中, 池化层不改变维度大小, ReLU 表示 ReLU 激活函数.

表 1 3D 卷积神经网络的结构

网络层名称	尺寸大小	维度
3D 卷积+ReLU	7×7×7	64
3D 最大池化	3×3×3	
3D 卷积+ReLU	3×3×3	128
3D 卷积+ReLU	3×3×3	128
3D 卷积+ReLU	3×3×3	256
3D 卷积+ReLU	3×3×3	256
3D 平均池化	3×3×3	

将提取到的视频局部和全局时域特征进行拼接; 再使用 1×1 卷积层降维, 得到融合局部和全局的时域特征 F_t , 最后经过 GAP 和 GSP 整合, 得到最终的时域特征 f_t . 该过程的公式为

$$F_t = \text{Conv1}(\text{Cat}(\text{VST}(M), \text{Conv3D}(M^G))),$$

$$f_t = \text{Cat}(\text{GAP}(F_t), \text{GSP}(F_t)).$$

其中, $\text{VST}(\cdot)$ 表示 VST 模型提取操作; $\text{Conv1}(\cdot)$ 表示 1×1 卷积操作; $\text{Conv3D}(\cdot)$ 表示 3D 卷积提取操作; M 表示输入视频帧序列; M^G 表示输入视频的帧差序列.

2.3 时空特征融合增强模块

通过时域特征提取模块和空域特征提取模块得到时域特征和空域特征后, 需要将 2 类特征进行融合, 用于最终的视频质量分数预测. 已有工作中, 常用的特征融合方法是直接将两者进行拼接. 但是研究表明^[35], 时域特征和空域特征对视频的质量影响程度是不同的, 因此单纯使用拼接的方式进行特征融合无法对 2 种特征信息的重要程度进行合理利用, 导致最终对视频的质量预测不够准确. 为了解决这个问题, 本文提出时空特征融合增强模块, 由自注意力子模块、残差连接、层归一化(layer normalization, LN)和前馈神经网络(feed forward, FFN)子模块组成, 其结构如图 4 所示. 通过自注意力机制学习 2 种不同的特征, 捕捉到时域

和空域特征之间的复杂关系, 并根据两者的重要程度分配权重并进行融合; 使用残差连接与空域特征相连的目的是保留更多的空域信息, 因为空域信息比时域信息对于最终的视频质量评价更重要^[35]; 层归一化和前馈神经网络子模块用于整合特征, 以得到更准确的质量分数。

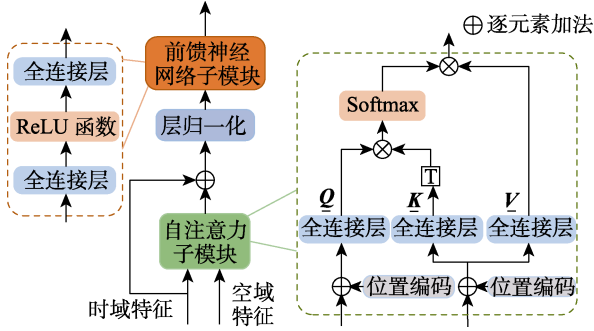


图 4 时空特征增强融合模块结构

对于提取的时域特征 $f_t \in \mathbb{R}^{0.5 \times L \times C}$ 和空域特征 $f_s \in \mathbb{R}^{L \times C}$, 其中, L 表示视频帧数, 先分别加上对应的可学习的位置编码(positional encoding, PE), 然后送入自注意力子模块进行运算, 并与空域特征 f_s 进行残差连接; 经过 LN, 再送入 FFN 子模块得到最终特征 $f_{fin} \in \mathbb{R}^{L \times C}$. 该过程的公式为

$$\begin{aligned} \underline{Q} &= FC(f_s + PE_s), \\ \underline{K}, \underline{V} &= FC(f_t + PE_t), \\ f_{st} &= \text{Softmax}(\underline{Q} \circ \underline{K}^T) \circ \underline{V}, \\ f_{fin} &= \text{FFN}(\text{LN}(f_{st} + f_s)). \end{aligned}$$

其中, $FC(\cdot)$ 表示全连接层(full connect, FC); PE_s 和 PE_t 表示可学习的 PE; $\text{FFN}(\cdot)$ 表示 FFN 子模块; $\text{LN}(\cdot)$ 表示 LN 操作。

2.4 质量回归和损失函数

将时空融合增强模块得到的最终特征 f_{fin} 送入 FC, 得到输入视频序列中每帧的质量分数. 在最后整合完整视频的质量分数时, 虽然已有一些整合完整视频的质量分数的方法^[18,36], 但是文献[24]的结果表明, 平均运算比这些方法更加有效. 因此本文对所有帧的质量分数进行平均运算, 得到预测的整个视频质量分数为

$$S = \frac{1}{N} \sum_{i=1}^N q_i.$$

其中, q_i 表示第 i 帧的质量分数; N 表示整个视频的帧数。

本文采用 L_1 损失作为损失函数, 公式为

$$L = \frac{1}{m} \sum_{i=1}^m (|S_i - \hat{S}_i|).$$

其中, m 表示批处理中视频的数量; S_i 和 $\hat{S}_i$ 分别表示批处理中第 i 个视频的本文方法预测的质量分数和真实质量分数。

3 实验及结果分析

3.1 数据集和实验细节

为了验证本文方法的有效性, 在 2 个主流的无参考视频质量评价数据集 KoNViD-1k^[16]和 YouTube-UGC^[17]上进行实验. 从大型公共视频数据库 YFCC100M^[37]中, 通过模糊、色彩、对比度、空域、时域和视频质量 6 个维度筛选出 KoNViD-1k 数据集, 其包含 1 200 个真实世界拍摄的视频, 每个视频的分辨率为 540 像素, 持续时间为 8 s, 主观质量评分范围在[1.22, 4.64]. YouTube-UGC 数据集则包含 YouTube 官网上 1 500 个时长 20 s 的视频, 视频内容包括游戏、体育、动画等 15 个类别, 视频分辨率从 360 像素到 4 000 像素不等, 质量评分范围在[1.242, 4.698].

本文方法基于 PyTorch 1.12.1 深度学习框架, 使用一张 Tesla P100-PCIE-16 GB GPU 进行训练和测试. 采取常见的训练策略: 对每个数据集随机选取 80%的视频作为训练集, 其余 20%作为测试集, 且这 2 部分没有重叠, 重复 5 次取平均值作为最终模型的性能评估指标. 由于 YouTube-UGC 数据集的视频分辨率不统一, 因此本文在时域特征提取模块和空域特征提取模块中的训练批次大小为 1, 而在时空特征融合模块中设置为 12; 为了与 YouTube-UGC 数据集统一, 本文在 KoNViD-1k 数据集上采用相同的批次大小设置。

本文网络的训练分为 3 个阶段: 对空域特征提取模块进行训练, 对时域特征提取模块进行训练, 对时空特征融合增强模块以及质量回归部分进行训练. 在对 2 个特征提取模块进行训练时, 使用 FC 得到视频序列每帧的质量分数, 取其平均值作为视频的质量分数, 并使用 L_1 损失作为损失函数; 3 个阶段均采用 Lion 算法^[38]优化损失函数. 在 2 个特征提取模块中, 设置初始学习率为 0.000 05, 衰减周期为 20 轮, 衰减率为 0.8; 在时空特征融合增强模块中, 设置初始学习率为 0.000 1, 衰减周期为 5 轮, 衰减率为 0.8。

第 1 阶段. 对空域特征提取模块进行训练. 由

于视频本身存在各种失真情况,而如果对视频进行裁剪或者重设大小会引入其他的失真影响现有的视频质量,因此本文方法没有对原始视频进行裁剪或者放缩到固定分辨率,而是使用原始视频作为输入.受到 GPU 设备的限制,本文将分辨率大于 1 080 像素的视频的宽度和高度均降为原视频的一半.与已有工作^[14]相同,网络先经过 KonIQ-10k 数据集进行预训练,将数据集按照 80%和 20%分为训练集和测试集,训练好的模型最终在测试集上的斯皮尔曼等级相关系数(Spearman rank order correlation coefficient, SROCC)指标为 0.930 2;然后在无参考视频质量数据集上进行训练.由于完整的视频中含有大量的视频帧,因此在训练空域特征提取模块时对每个视频进行均匀采样,采样频率为 f ,将其设置为 4;之后,保存模型并对视频中的所有视频帧提取空域特征.训练完成后, α 和 β 值分别为 0.4898 和 0.4928,因为卷积提取内容主体的局部信息而自注意力提取的是全局信息,而全局信息的重要性略低于主体的局部信息,因此 α 值略低于 β .

第 2 阶段.对时域特征提取模块进行训练.相对于空域信息,时域信息对分辨率的敏感性较低,为了加快训练,将视频的分辨率统一重设大小为 512×512 像素.经过训练之后保存模型并提取视频的时域特征.由于 VST 模型会将视频帧的长度下降一半,因此本文使用的帧差法是对每 2 个连续视频帧进行差值计算,且每帧不重复使用(若视频帧数为奇数则去除最后一帧),使得卷积提取的时域特征长度为原来的一半,与 VST 所提取的特征长度一致.

第 3 阶段.对时空特征融合增强模块以及质量回归部分进行训练.将前 2 个阶段训练完成的模型提取的空域特征和时域特征送入时空特征融合增强模块以及质量回归部分进行训练,得到最终的实验结果.

3.2 方法对比

为了验证本文方法的有效性,使用皮尔逊线性相关系数(Pearson linear correlation coefficient, PLCC)和 SROCC 作为性能评价指标,两者的指标值越接近 1,表示效果越好.将本文方法与具有代表性的 13 个无参考视频质量评价方法进行对比,包括 TLVQM^[5],HVS-5M^[6],VIDEVAL^[13],RAPIQUE^[15],VSFA^[18],PVQ^[23],SimpleVQA^[24],STDAM^[25],BVQA-2022^[26],MD-VQA^[27],BVQI^[39],CoINVQ^[40]和 CONVIQT^[41].表 2 所示为不同方法

与本文方法在 KoNVid-1k 和 YouTube-UGC 数据集上的实验结果对比.可以看出,本文方法在 2 个数据集上均取得了最优效果:在 KoNVid-1k 数据集上,与其中最好的 SimpleVQA^[24]方法相比,该方法在 SROCC 和 PLCC 上分别提升 0.022 和 0.016;在 YouTube-UGC 数据集上,该方法的 SROCC 和 PLCC 也比其中最好的 HVS-5M^[6]和 SimpleVQA^[24]方法更高.实验结果表明,本文方法对视频质量特征具有较强的感知能力,相较对比方法,能够获得与主观感知更一致的评价结果.

表 2 不同方法在 2 个数据集上的对比结果

方法	KoNVid-1k		YouTube-UGC	
	SROCC	PLCC	SROCC	PLCC
VSFA ^[18]	0.755	0.744	0.724	0.743
CoINVQ ^[40]	0.767	0.764	0.816	0.802
TLVQM ^[5]	0.773	0.769	0.657	0.647
VIDEVAL ^[13]	0.783	0.780	0.779	0.773
PVQ ^[23]	0.791	0.786		
RAPIQUE ^[15]	0.803	0.817	0.759	0.768
BVQI ^[39]	0.833	0.830	0.816	0.804
BVQA-2022 ^[26]	0.835	0.834	0.825	0.818
STDAM ^[25]	0.845	0.842	0.834	0.830
CONVIQT ^[41]	0.851	0.849	0.832	0.822
MD-VQA ^[27]	0.851	0.853		
HVS-5M ^[6]	0.853	0.856	<u>0.852</u>	0.845
SimpleVQA ^[24]	<u>0.856</u>	<u>0.860</u>	0.847	<u>0.856</u>
本文	0.878	0.876	0.861	0.868

注.粗体表示最优值,下划线表示次优值.

3.3 消融实验

分别对空域特征提取模块、时域特征提取模块和时空特征融合增强模块进行消融实验,验证本文方法中各个模块的有效性.在 YouTube-UGC 数据集上进行 4 个消融实验:单独使用时域特征提取模块、单独使用空域特征提取模块、采用通道拼接的方式融合时域特征和空域特征,以及本文方法.其中,在只有时域特征提取模块的消融实验中,进一步对 2 个分支进行消融实验,包括只有 VST,只有 3D 卷积和同时使用 VST 和 3D 卷积,结果如表 3 所示.可以看出,单独使用时域特征的效果不如单独使用空域特征的效果;联合使用 VST 和 3D 卷积能显著地提升时域特征提取模块的性能;与单纯使用空域特征相比,已有工作中常用的使用通道连接将时域特征和空域特征进行融合的拼接方法在 SROCC 和 PLCC 上分别提升 0.008 和 0.010;而

使用本文提出的时空特征融合增强模块在 SROCC 和 PLCC 上分别提升 0.026 和 0.031, 比使用通道拼接方法分别提升 225% 和 210%。消融实验结果充分表明了本文提出的时空融合增强模块的有效性。

表 3 在 YouTube-UGC 数据集上的消融实验对比

空域	时域	融合	SROCC	PLCC
√	×	×	0.835	0.837
×	VST	×	0.755	0.765
×	3D 卷积	×	0.588	0.580
×	VST 和 3D 卷积	×	0.797	0.802
√	√	拼接	0.843	0.847
√	√	√	0.861	0.868

注. 粗体表示最优值。

本文对 ResNet50 的优越性进行验证. 将其替换为 FasterNet^[42], 经过实验发现, 使用 FasterNet 的效果并不理想: 在图像质量评价数据集预训练时, 其 SROCC 和 PLCC 仅分别为 0.700 8 和 0.712 4, 而采用 ResNet50 时 SROCC 和 PLCC 分别为 0.930 2

和 0.939 8; 模型预训练后, 在视频质量评价数据集上进行空域特征提取模块的训练时, 其 SROCC 和 PLCC 更是分别低至 0.582 和 0.571, 而采用 ResNet50 时 SROCC 和 PLCC 分别为 0.835 和 0.837. 因此, 本文选用 ResNet50 作为空域特征提取网络。

本文还验证了特征感知增强模块的卷积分支和自注意力分支的重要性, 并对常见的通道注意力和卷积注意力模块^[43]进行了消融实验; 其中, 卷积注意力模块结合了空域注意力和通道注意力 2 种注意力机制, 结果如表 4 所示. 不同方法的特征热力图如图 5 所示。

表 4 不同注意力模块在空域上的消融实验对比

方法	SROCC	PLCC
卷积	0.826	0.814
自注意力	0.822	0.823
卷积注意力模块	0.823	0.817
通道注意力	0.825	0.822
特征感知增强模块	0.835	0.837

注. 粗体表示最优值。

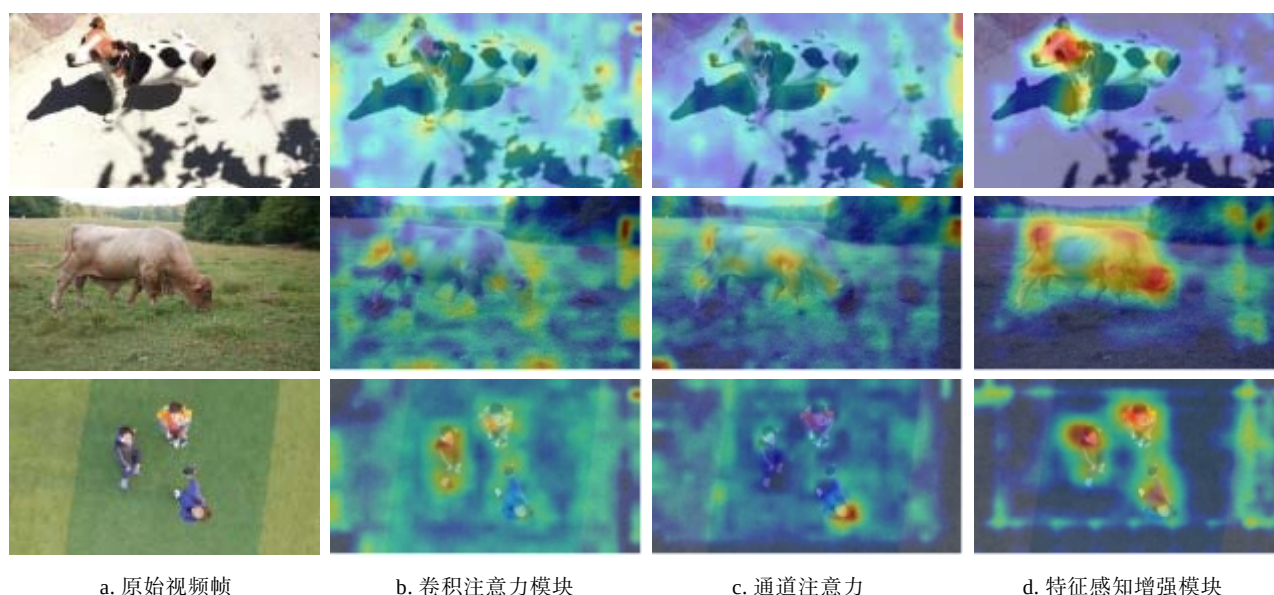


图 5 不同方法的特征热力图

从表 4 可以看出, 本文提出的特征感知增强模块在空域特征提取上取得了最好的效果, 优于单一的卷积和自注意力方法, 同时也优于通道注意力和卷积注意力模块; 卷积注意力模块的结果不如通道注意力, 这是因为前者过度关注全局信息, 导致模型降低了对影响图像质量的关键区域的注意力, 从而影响了模型的效果. 从图 5 可以看出, 卷积注意力模块和通道注意力模块过多地关注全

局信息, 而缺乏对于视频内容主体区域的聚焦(其中, 卷积注意力模块关注的全局信息多于通道注意力模块); 而本文的特征感知增强模块能同时合理地视频的局部和全局信息进行关注。

进一步, 对空域特征提取模块的训练阶段的采样频率进行消融实验, 验证了不同采样频率对空域特征的结果影响, 结果如表 5 所示. 可以看出, 采样频率为 1, 2, 4 时, 模型的性能几乎一致; 与采

样频率 1 和采样频率 2 相比, 选取采样频率 4 时节约了 75%和 50%的资源, 而继续增加采样频率会导致模型性能明显下降. 因此, 本文选择 4 作为空域特征提取模块训练时的采样频率.

表 5 不同采样频率对空域上的消融实验结果

采样频率	SROCC	PLCC
1	0.836	0.838
2	0.837	0.835
4	0.835	0.837
6	0.826	0.819
8	0.820	0.821

在池化方法的选择上, 本文进行了 GAP、全局最大池化(global max pooling, GMP)和 GSP 这 3 种方法的消融实验, 结果如表 6 所示. 可以看出, GAP 与 GSP 的组合效果最好, GMP 的效果最差, 且其他结合 GMP 的组合效果都不如原有方法. 原因是 GMP 仅关注图像中的局部最大值, 忽略了其他图像信息的影响, 并且在与其他方法组合时使用了过多冗余特征信息, 在后续学习过程中难以有效地学习和利用这些信息; 而 GAP 能够捕捉到区域内的整体信息, GSP 则关注区域内的数据波动, 可以捕捉到特征的离散程度, 两者结合可以使模型更好地捕捉特征多样性, 从而提高模型的表达能力.

表 6 不同池化方法在空域上的消融实验结果

方法	SROCC	PLCC
GAP	0.831	0.833
GSP	0.828	0.820
GMP	0.816	0.812
GAP+GMP	0.822	0.824
GSP+GMP	0.820	0.817
GAP+GSP	0.835	0.837
GAP+GSP+GMP	0.823	0.828

注. 粗体表示最优值.

4 结 语

针对现有视频质量评价方法中未能同时对视频提取局部和全局感知特性, 以及时空特征融合利用不足的问题, 本文提出卷积与自注意力结合和时空融合在无参考视频质量评价方法. 通过结合卷积和自注意力机制的特征感知增强模块, 在视频的空域维度提取富有局部和全局信息的空域

特征; 在时域维度, 使用 VST 和 3D 卷积神经网络提取全局和局部的时域特征; 通过时空特征融合增强模块将时域和空域特征进行融合并增强, 最后对视频质量进行预测. 在 KoNVid-1k 和 YouTube-UGC 数据集上的实验结果表明, 本文方法比现有方法具有更出色的表现; 消融实验也表明了该方法中每个模块的有效性. 尽管如此, 本文所提模型结构也更为复杂, 增加了计算开销, 未来我们将探索更精简、更高效的网络模型.

参考文献(References):

- [1] Duanmu Z F, Zeng K, Ma K D, *et al.* A quality-of-experience index for streaming video[J]. *IEEE Journal of Selected Topics in Signal Processing*, 2017, 11(1): 154-166
- [2] Yan Jiebin, Fang Yuming, Liu Xuelin, *et al.* A survey on recent advances in video quality assessment[J]. *Chinese Journal of Computers*, 2023, 46(10): 2196-2224(in Chinese)
(鄢杰斌, 方玉明, 刘学林, 等. 视频质量评价研究综述[J]. *计算机学报*, 2023, 46(10): 2196-2224)
- [3] Liao L, Xu K M, Wu H N, *et al.* Exploring the effectiveness of video perceptual representation in blind video quality assessment[C] // *Proceedings of the 30th ACM International Conference on Multimedia*. New York: ACM Press, 2022: 837-846
- [4] Mittal A, Saad M A, Bovik A C. A completely blind video integrity oracle[J]. *IEEE Transactions on Image Processing*, 2016, 25(1): 289-300
- [5] Korhonen J. Two-level approach for no-reference consumer video quality assessment[J]. *IEEE Transactions on Image Processing*, 2019, 28(12): 5923-5938
- [6] Golestaneh S A, Dadsetan S, Kitani K M. No-reference image quality assessment via transformers, relative ranking, and self-consistency[C] // *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*. Los Alamitos: IEEE Computer Society Press, 2022: 3989-3999
- [7] Deng J, Dong W, Socher R, *et al.* ImageNet: a large-scale hierarchical image database[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2009: 248-255
- [8] He K M, Zhang X Y, Ren S Q, *et al.* Deep residual learning for image recognition[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2016: 770-778
- [9] Liu Z, Ning J, Cao Y, *et al.* Video swin transformer[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2022: 3192-3201
- [10] Zou Liangtao, Jiang Gangyi, Yu Mei, *et al.* No-reference image quality assessment of high dynamic range image based on tensor domain perceptual features[J]. *Journal of Computer-Aided Design & Computer Graphics*, 2018, 30(10): 1850-1858(in Chinese)
(邹良涛, 蒋刚毅, 郁梅, 等. 基于张量域感知特征的无参考高动态范围图像质量评价[J]. *计算机辅助设计与图形学学*

- 报, 2018, 30(10): 1850-1858)
- [11] 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
 - [12] 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
 - [13] Tu Z Z, Wang Y L, Birkbeck N, *et al.* UGC-VQA: benchmarking blind video quality assessment for user generated content[J]. *IEEE Transactions on Image Processing*, 2021, 30: 4449-4464
 - [14] Korhonen J, Su Y C, You J Y. Blind natural video quality prediction via statistical temporal features and deep spatial features[C] // *Proceedings of the 28th ACM International Conference on Multimedia*. New York: ACM Press, 2020: 3311-3319
 - [15] Tu Z Z, Yu X X, Wang Y L, *et al.* RAPIQUE: rapid and accurate video quality prediction of user generated content[J]. *IEEE Open Journal of Signal Processing*, 2021, 2: 425-440
 - [16] Hosu V, Hahn F, Jenadeleh M, *et al.* The konstanz natural video database(KoNViD-1k)[C] // *Proceedings of the 9th International Conference on Quality of Multimedia Experience*. Los Alamitos: IEEE Computer Society Press, 2017: 1-6
 - [17] Wang Y L, Inguva S, Adsumilli B. YouTube UGC dataset for video compression research[C] // *Proceedings of the 21st IEEE International Workshop on Multimedia Signal Processing*. Los Alamitos: IEEE Computer Society Press, 2019: 1-5
 - [18] Li D Q, Jiang T T, Jiang M. Quality assessment of in-the-wild videos[C] // *Proceedings of the 27th ACM International Conference on Multimedia*. New York: ACM Press, 2019: 2351-2359
 - [19] 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
 - [20] Shahreza H O, Amini A, Behroozi H. No-reference video quality assessment using recurrent neural networks[C] // *Proceedings of the 5th Iranian Conference on Signal Processing and Intelligent Systems*. Los Alamitos: IEEE Computer Society Press, 2019: 1-5
 - [21] Chen P F, Li L D, Ma L, *et al.* RIRNet: recurrent-in-recurrent network for video quality assessment[C] // *Proceedings of the 28th ACM International Conference on Multimedia*. New York: ACM Press, 2020: 834-842
 - [22] Hou R, Zhao Y H, Hu Y, *et al.* No-reference video quality evaluation by a deep transfer CNN architecture[J]. *Signal Processing: Image Communication*, 2020, 83: 115782
 - [23] Ying Z Q, Mandal M, Ghadiyaram D, *et al.* Patch-VQ: 'patching up' the video quality problem[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2021: 14014-14024
 - [24] Sun W, Min X K, 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*. New York: ACM Press, 2022: 856-865
 - [25] Xu J H, Li J, Zhou X G, *et al.* Perceptual quality assessment of internet videos[C] // *Proceedings of the 29th ACM International Conference on Multimedia*. New York: ACM Press, 2021: 1248-1257
 - [26] Li B W, Zhang W X, Tian M, *et al.* Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022, 32(9): 5944-5958
 - [27] Zhang Z C, Wu W, Sun W, *et al.* MD-VQA: multi-dimensional quality assessment for UGC live videos[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2023: 1746-1755
 - [28] Zhang A X, Wang Y G, Tang W X, *et al.* HVS revisited: a comprehensive video quality assessment framework[OL]. [2023-12-18]. <https://arxiv.org/abs/2210.04158>
 - [29] Kay W, Carreira J, Simonyan K, *et al.* The kinetics human action video dataset[OL]. [2023-12-18]. <https://arxiv.org/abs/1705.06950>
 - [30] Wu Jun, Cui Yue, Zhao Xuemei, *et al.* SSA-PointNet++: aspace self-attention CNN for the semantic segmentation of 3D point cloud[J]. *Journal of Computer-Aided Design & Computer Graphics*, 2022, 34(3): 437-448(in Chinese)
(吴军, 崔玥, 赵雪梅, 等. SSA-PointNet++: 空间自注意力机制下的 3D 点云语义分割网络[J]. *计算机辅助设计与图形学学报*, 2022, 34(3): 437-448)
 - [31] Guo Danqing, Fu Ying, Zhu Ye, *et al.* Semantic segmentation of remote sensing image via self-attention-based multi-scale feature fusion[J]. *Journal of Computer-Aided Design & Computer Graphics*, 2023, 35(8): 1259-1268(in Chinese)
(郭丹青, 符颖, 朱烨, 等. 自注意力多尺度特征融合的遥感图像语义分割算法[J]. *计算机辅助设计与图形学学报*, 2023, 35(8): 1259-1268)
 - [32] Wen Yuanbo, Gao Tao, Chen Ting, *et al.* Frequency-guided dual sparse self-attention algorithm for single image deraining[J]. *Acta Electronica Sinica*, 2023, 51(10): 2812-2820(in Chinese)
(文渊博, 高涛, 陈婷, 等. 频率引导的双稀疏自注意力单图像去雨算法[J]. *电子学报*, 2023, 51(10): 2812-2820)
 - [33] Huan Ying, He Zifen, Yang Hongkuan, *et al.* Multi-scale segmentation of episodic video instance through polarized self-attention manipulation[J]. *Chinese Journal of Computers*, 2022, 45(12): 2605-2618(in Chinese)
(黄滢, 何自芬, 杨宏宽, 等. 极化自注意力调控的情景式视频实例多尺度分割[J]. *计算机学报*, 2022, 45(12): 2605-2618)
 - [34] Hara K, Kataoka H, Satoh Y. Learning spatio-temporal features with 3D residual networks for action recognition[C] // *Proceedings of the IEEE International Conference on Computer Vision Workshops*. Los Alamitos: IEEE Computer Society Press, 2017: 3154-3160
 - [35] Huynh-Thu Q, Ghanbari M. Modelling of spatio-temporal interaction for video quality assessment[J]. *Signal Processing: Image Communication*, 2010, 25(7): 535-546
 - [36] Tu Z Z, Chen C J, Chen L H, *et al.* A comparative evaluation of temporal pooling methods for blind video quality assessment[C] // *Proceedings of the IEEE International Conference on Image Processing*. Los Alamitos: IEEE Computer Society Press, 2020: 141-145
 - [37] Thomee B, Shamma D A, Friedland G, *et al.* YFCC100M: the new data in multimedia research[J]. *Communications of the ACM*, 2016, 59(2): 64-73

- [38] Chen X N, Liang C, Huang D, *et al.* Symbolic discovery of optimization algorithms[C] //Proceedings of the 37th International Conference on Neural Information Processing Systems. Red Hook:Curran Associates Inc., 2023: Article No.2140
- [39] Wu H N, Liao L, Wang A N, *et al.* Towards robust text-prompted semantic criterion for in-the-wild video quality assessment[OL]. [2023-12-18]. <https://arxiv.org/abs/2304.14672>
- [40] Wang Y L, Ke J J, Talebi H, *et al.* Rich features for perceptual quality assessment ofUGC videos[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 13430-13439
- [41] Madhusudana P C, Birkbeck N, Wang Y L, *et al.* CONVIQT: contrastive video quality estimator[J]. IEEE Transactions on Image Processing, 2023, 32: 5138-5152
- [42] Chen J R, Kao S H, He H, *et al.* Run, Don't walk: chasing higher flops for faster neural networks[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2023: 12021-12031
- [43] Woo S, Park J, Lee J Y, *et al.* CBAM: convolutional block attention module[C] //Proceedings of the 15th European Conference on Computer Vision. Heidelberg: Springer, 2018: 3-19