计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,xxxx,xx(x):x−x [doi: 10.15888/j.cnki.csa.00xxxx]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041.

基于自监督学习与多尺度时空特征融合的视频质量评估①

于莉[1], 王思拓[1], 陈亚当[1], 高攀[2], 孙玉宝[1]

[1](南京信息工程大学 计算机学院、网络空间安全学院, 南京 210044)
[2](南京航空航天大学 计算机科学与技术学院, 南京 211106)
通讯作者: 于莉, E-mail: li.yu@nuist.edu.cn

摘 要: 面对视频质量评估领域标记数据不足的问题, 研究者开始转向自监督学习方法, 旨在借助大量未标记数据以学习视频质量评估模型. 然而现有自监督学习方法主要聚焦于视频的失真类型和视频内容信息, 忽略了视频随时间变化的动态信息和时空特征, 这导致在复杂动态场景下的评估效果不尽人意. 针对上述问题, 提出了一种新的自监督学习方法, 通过播放速度预测作为预训练的辅助任务, 使模型能更好地捕捉视频的动态变化和时空特征, 并结合失真类型预测和对比学习, 增强模型对视频质量差异的敏感性学习. 同时, 为了更全面捕捉视频的时空特征, 进一步设计了多尺度时空特征提取模块等以加强模型的时空建模能力. 实验结果显示, 所提方法在 LIVE、CSIQ 以及 LIVE-VQC 数据集上, 性能显著优于现有的基于自监督学习的方法, 在 LIVE-VQC 数据集上, 本方法在 PLCC 指标上平均提升 7.90%, 最高提升 17.70%. 同样, 在 KoNVid-1k 数据集上也展现了相当的竞争力. 这些结果表明, 提出的自监督学习框架有效增强了视频质量评估模型的动态特征捕捉能力, 并在处理复杂动态视频中显示出独特优势.

关键词: 视频质量评估; 自监督学习; 多任务学习; 播放速度预测; 多尺度

Self-Supervised Learning and Multi-Scale Spatio-Temporal Feature Fusion Based Video Quality Assessment

YU Li[1], WANG Si-Tuo[1], CHEN Ya-Dang[1], GAO Pan[2], SUN Yu-Bao[1]

[1](School of Computer Science, School of Cyber Science and Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China)
[2](College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

Abstract: Faced with the issue of insufficient labeled data in the field of video quality assessment, researchers have turned to self-supervised learning methods, leveraging large amounts of unlabeled data to train video quality assessment models. However, existing self-supervised methods primarily focus on video distortion types and content information, ne-glecting dynamic information and spatiotemporal features of videos over time. This leads to suboptimal performance in complex dynamic scenes. To address these issues, a new self-supervised learning method is proposed, using playback speed prediction as an auxiliary pretraining task. This enables the model to better capture dynamic changes and spatiotemporal features of videos. Combined with distortion type prediction and contrastive learning, the model enhances its sensitivity to video quality differences. Additionally, to more comprehensively capture the spatiotemporal features of the video, a multi-scale spatiotemporal feature extraction module was further designed to enhance the model's spatiotemporal modeling capability. Experimental results demonstrate that the proposed method significantly outperforms existing self-supervised learning-based approaches on the LIVE, CSIQ, and LIVE-VQC datasets. Specifically, on the LIVE-VQC

① 基金项目:国家自然科学基金(62002172,62276139,U2001211)
收稿时间:xxxx-xx-xx;收到修改稿时间:xxxx-xx-xx

dataset, our method achieves an average improvement of 7.90% in the PLCC metric, with a maximum improvement of 17.70%. Similarly, it also shows strong competitiveness on the KoNVid-1k dataset. These results indicate that the proposed self-supervised learning framework effectively enhances the model's ability to capture dynamic features and exhibits unique advantages in evaluating complex dynamic videos.

Key words: video quality assessment; self-supervised learning; multi-task learning; playback speed prediction; multi-scale

随着社交媒体和视频分享平台的飞速发展, 用户生成的视频内容呈现出爆炸式增长. 这些视频的质量受到多方面因素的制约, 包括存储、传输和播放过程中可能引发的质量退化等[1]. 因此, 对视频内容进行质量评估已成为计算机视觉领域的一个重要任务, 以确保用户能够获得良好的观看体验. 视频质量评估(Video Quality As-sessment, VQA)可分为主观和客观两种方法. 主观评估基于人类的观察和感知, 结果较为准确, 但因其高昂的成本、实验环境要求和无法实时评估等局限性, 使其在实际应用中受限. 因此, 该文主要考虑使用客观质量评估. 这种方法不需要人类评估, 而是使用算法自动评估视频质量. 客观质量评估算法按照对参考视频的使用程度可以分为三类：全参考视频质量评估(Full-reference VQA, FR-VQA)[2,3]、半参考视频质量评估(Reduced-reference VQA, RR-VQA)[4]和无参考视频质量评估(No-reference VQA, NR-VQA)[5-15]. 目前, 许多视频由于网络传输或者用户拍摄水平等因素的限制, 导致其无失真的原始参考视频几乎无法获取, 因此不依赖原始参考视频, 直接基于失真视频进行质量评价的无参考视频质量评估显的尤为重要, 它也是目前研究人员研究的重点.

在传统的无参考视频质量评估中, 常见的做法是利用有监督学习方法自动提取视频特征. 这些方法需要依赖人工标记的数据集, 通过在标记数据上进行模型训练, 从中自动提取视频特征以评估不同视频的质量. 然而, 这种方法在实际应用中存在明显的局限性, 主要问题在于获取高质量的标记数据既费时又昂贵, 且难以全面覆盖多样化的视频质量问题. 因此, 自监督学习方法因其能够利用大量未标记的数据而日益受到关注. 自监督学习的核心思想在于设定辅助任务生成伪训练标签, 这使得模型能在无需人工标注的情况下进行训练. 这种方法的优势在于, 研究人员可以使用大量的公开视频数据对模型进行训练和优化. 近年来, 基于自监督学习的VQA的方法[17-20]取得了一定的进展. 文献[17]提出使用对比学习的方式对视频质量评价任务进行自监督预训练, 文献[19]的模型则受到了自监督图像质量评估模型 CONTRIQUE[27]的启发, 将失真类型识别和失真程度确定作为其训练的辅助任务来预训练 VQA 模型. 文献[20]通过结合失真程度、失真类型和帧率的自监督信息来捕捉视频的特征. 然而, 这些方法忽略了视频的时空特性和动态特征, 这限制了它们在处理复杂动态场景时的表现. 尽管一些工作[18,20]尝试整合更多信息以改善这种局限性, 但仍然难以全面捕捉视频的所有关键动态特征.

在此背景之下, 本项研究创新地提出了基于自监督学习与多尺度时空特征融合的视频质量评估方法. 首次将播放速度这一新维度作为自监督信号, 以探索视频的深层时空信息. 具体而言, 该文通过对视频帧间隔采样以模拟不同播放速度, 并自动生成对应的播放速度标签. 通过构造不同播放速度的视频丰富了视频内容随时间变化的细节, 强制模型识别和理解视频在不同速度下的表现, 极大地丰富了模型对视频时空结构的理解, 使其能够识别复杂的时空特性. 相较而言, 传统基于帧率预测的方法主要关注视频的整体流畅性而非内容随时间的具体变化模式, 因而无法充分捕获细微的动态变化. 除此之外, 该文还通过融合对比学习策略以进一步强化视频特征的表示能力. 通过结合播放速度预测任务和对比学习充分挖掘视频的时空特征, 突破了传统自监督学习仅关注于视频的失真类型、失真程度或帧率等单一属性的局限, 为视频质量评估任务提供了更加全面、可靠的视频特征表示. 同时, 为了提高VQA模型在感知范围和时空建模方面的性能, 该文设计了多尺度时空特征提取模块提取时空特征, 并通过 SimAM 注意力机制[21]提高模型对关键特征的关注度. 此外, 该文还引入了 Swin Transformer[22]进一步提升空间特征的编码效率和准确

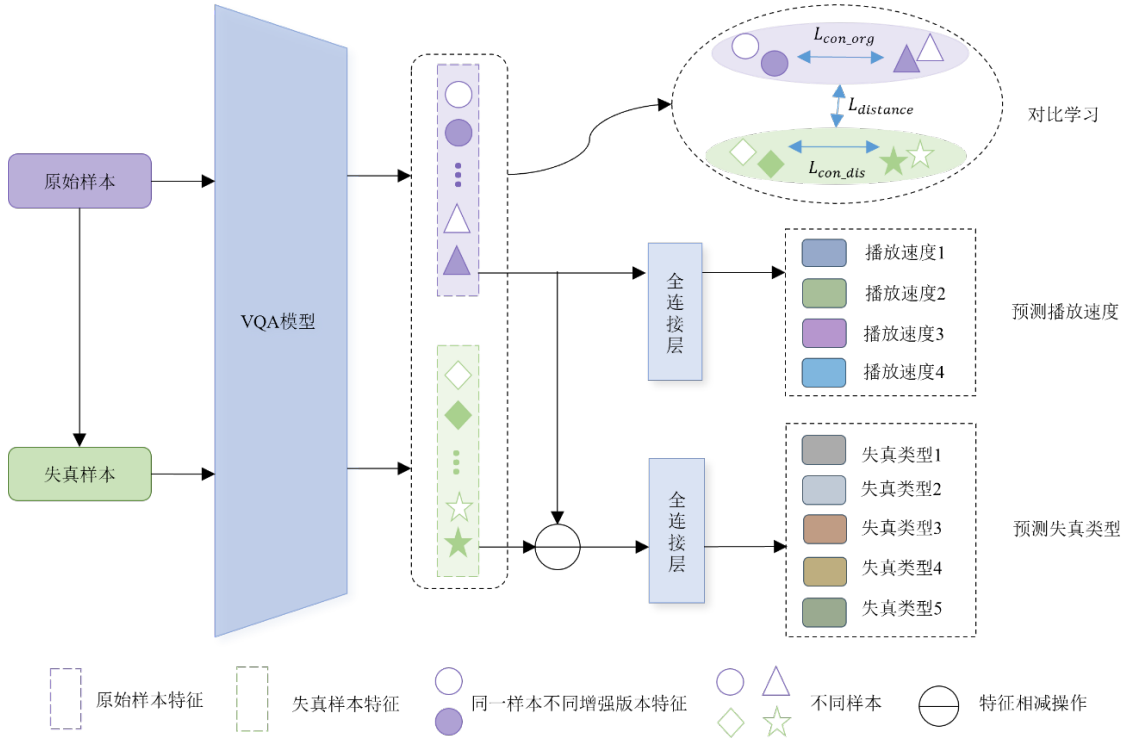


图 1 该文所提自监督方法部分的整体框架图

性，以全面提高模型的时空特征处理能力，从而提高视频质量评估的准确性和鲁棒性. 在两个真实失真数据集和两个合成失真数据集上的实验结果验证了该文自监督学习策略的有效性，也展示了该方法在动态视频质量评价上的潜力.

1 该文方法

该文所提方法的整体框架如图 1 所示. 首先将原始样本和失真样本分别输入到VQA模型，然后通过三个自监督辅助任务对VQA模型进行预训练. 该方法旨在通过自监督学习提高 VQA 模型在捕捉视频质量相关特征方面的能力，以便更准确地判断视频内容的质量. 为了实现这个目的，该文首先构建了一个大规模的数据集，并生成自注意标签，作为自监督学习的基础. 在模型的自监督预训练阶段，VQA 模型通过预测播放速度、预测失真类型以及对比学习这三个自监督任务进行预训练. 自监督数据集的构建方法将在 1.1 节介绍，该文所使用的三个自监督任务将在 1.2 节中详细阐述. 继预训练之后，再使用公共 VQA 数据集对模型进行微调，使模型能够提供更准确的视频质量评分. VQA 模型将在第 2 节中进行详细说明.

1.1 自监督数据集准备

在视频质量评估任务中，获取大量精确标注的数据既昂贵又耗时，尤其是在手动标注视频的平均意见得分(Mean Opinion Score, MOS)时，人力和时间成本极高. 为解决这一问题，该研究采用自监督学习，通过自动生成标签以减少对人工标注的依赖. 首先，从YouTube-8M 数据库中筛选了 310 个视频，并对其应用 5 种失真处理，包括高斯模糊、对比度调整、H.264 压缩、运动模糊和高斯噪声. 这些失真类型将在模型预训练阶段作为自监督标签，用于失真预测任务

考虑到播放速度预测对深入探索视频深层时空信息的重要性，该研究通过视频帧间隔采样的方法自动生成播放速度标签以模拟不同的播放速度. 具体实现方式如下，首先定义一个变量 *sample_rate*，取值为{1、2、3、4}，然后从中随机取一个数字代表播放速度，如果我们选择的播放速度为 1，那么就代表是正常的播放速度，训练片段从原始样本中连续采样即可；如果 *sample_rate*>1，则训练片段每隔 *sample_rate* 帧采样一次. 播放速度标签通过 *sample_rate-1* 得到，将范围映射到[0, *max_sr*-1], *max_sr* 代表随机生成的播放速度的最大值. 这样，播放速度标签即表示播放速度用于预训练模型.

最终，将得到一个完整的自监督数据集用于预训

练VQA模型. 该数据集包含了失真类型和播放速度在内的标签, 提供了必要的监督信息, 为后续模型的有效训练提供了数据基础.

1.2 自监督辅助任务

该研究引入了一种创新的自监督学习框架, 使用播放速度预测、失真类型预测以及对比学习作为辅助任务. 具体地, 播放速度预测任务迫使模型注意视频内容随时间的变化, 提高模型捕捉视频动态变化和时空特征的能力, 失真类型预测和对比学习任务则加强了模型在不同失真条件下的特征区分能力, 增强了模型对视频差异的敏感性学习. 这三个任务各自解决特定问题的同时也相互补充, 共同构成一个强大的多任务自监督学习框架, 显著提升了VQA模型在各种场景下的性能和泛化能力.

1.2.1 预测失真类型

在视频质量评估任务中, 识别和预测失真类型是至关重要的, 因为它直接关联到视频质量的感知和评价. 虽然在原始视频上添加的失真类型是确定的, 但原始视频是从 YouTube-8M 中随机提取的, 可能存在未知失真. 因此, 为了避免未知失真影响失真预测的准确性, 该研究聚焦于原始视频$Video_{ori}$与其失真版本$Video_{dis}$之间的时空信息差异, 并利用这些差异以预测失真视频的失真类型, 从而提高VQA模型对失真类型的预测能力. 基于此, 该文使用暹罗网络$Siamese(.)$处理失真样本和原始样本. 该方法的优势在于可以利用原始样本和失真样本之间的差异信息Fea_{diff}预测失真类型, 使得模型能够较好地适应其他类型的失真, 而不仅仅依赖于单独的原始样本或失真样本的特征, 如公式(1)所示.

$$Fea_{diff} = Siamese(Video_{ori}) - Siamese(Video_{dis})$$
$$d' = FC(Fea_{diff}) \tag{1}$$

$Siamese(Video_{ori})$和$Siamese(Video_{dis})$代表通过暹罗网络提取的特征, 将得到的差异信息输入到全连接层中, 得到预测的失真类型的概率d'. 通过对差异特征的学习和分析, 可以使网络更好地理解失真对样本特征的影响, 从而提高失真类型的预测性能.

1.2.2 预测播放速度

该文选择播放速度预测作为预训练阶段的自监督任务, 旨在促进模型对视频内容信息及其时空特性的深入理解. 该文认为只有模型理解了视频内容并且学到了整体的时空特征表示才能够准确的预测播放速度. 由于播放速度预测任务不需要考虑原始样本和失真样本之间的差异. 因此, 在预测播放速度时, 该文仅使用原始样本$Video_{ori}$作为输入:

$$p' = \mathrm{FC_{pace}}\left(\mathrm{Siamese}(\mathrm{Video_{ori}})\right) \tag{2}$$

p'代表通过该文的模型预测得到的播放速度的概率, FC_{pace}代表用于预测播放速度的全连接网络.

预测失真类型和预测播放速度这两个辅助任务都采用交叉熵损失函数.

$$l_{dis} = -\sum_{i=0}^{N-1} d_i(\ln \frac{e^{\hat{d}_i}}{\sum_{j=1}^{N-1} e^{\hat{d}_j}}) \tag{3}$$

$$l_{pace} = -\sum_{i=0}^{M-1} p_i(\ln \frac{e^{\hat{p}_i}}{\sum_{j=1}^{M-1} e^{\hat{p}_j}}) \tag{4}$$

其中d是失真类型标签, $\hat{d}$是失真类型预测值, N 是所有失真类型的数量, p代表播放速度标签, $\hat{p}$是播放速度预测值, M 是所有播放速度类型的数量. l_{dis}是用于进行失真类型预测的损失函数, l_{pace}是用于进行播放速度预测的损失函数.

1.2.3 对比学习

为了充分地提取原始样本与失真样本中的信息, 并确保模型能够清晰地区分二者, 该文采用了一种结合对比学习和余弦相似度的方法. 通过对这两种类型的样本分别实施对比学习, 模型能够学习到各自的独特特征. 同时, 引入余弦相似度作为一种补充机制, 进一步增强模型对原始与失真样本之间差异的识别能力.

首先, 定义特征向量间的相似度度量. 在特征空间中, 常用的方法是使用点积计算相似度, 它同时考虑了特征向量的方向和长度信息, 这有助于捕捉特征向量的整体差异性.

$$sim(x, y) = x^T y \tag{5}$$

其中 x 和 y 代表两个不同的特征向量, 可以通过计算它们的点积从而衡量它们之间的相似性, 这一相似性可以用$sim(.)$函数表示.

考虑原始样本的对比学习, 该文采用数据增强技术, 如裁剪、翻转等, 获取原始样本的两个不同增强版本, 并将其定义为正样本对, 不同原始样本定义为负样本对, 其目标是使模型能够将原始样本的正样本对的特征拉近, 负样本对的特征拉远. 与先前工作[17]类似, 该研究采用了 InfoNCE 损失函数[23]以定义原始样本的损失, 该函数已被证明能有效地在表示学习中增强正样本之间的相似性, 并将负样本分开, 具体的损失函数定义如下:

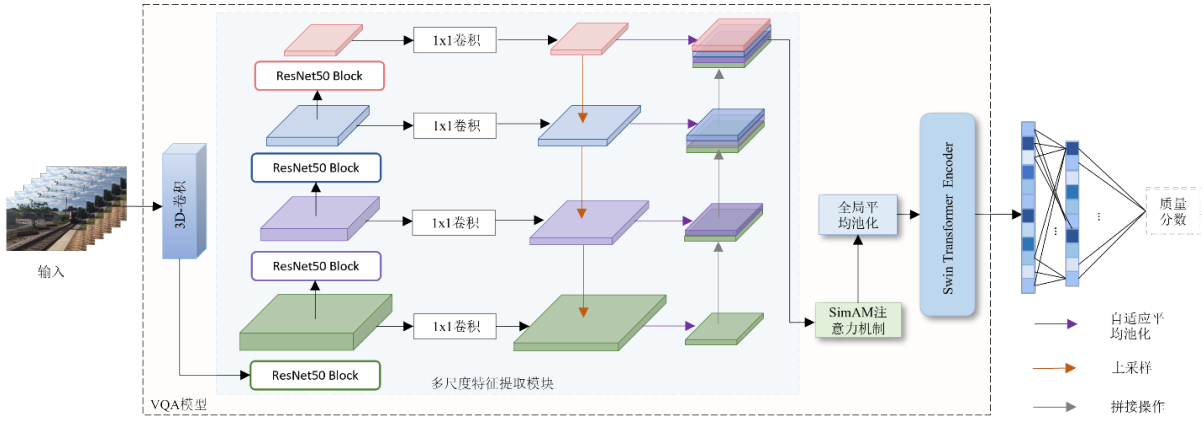


图 2 该文所提 VQA 模型框架图

$$l_{\text{con\_org}} = -\ln \frac{\exp(\text{sim}(z_o, z_{o'})/T)}{\sum_{k=1}^{2N} 1_{[k\neq o]} \exp(\text{sim}(z_o, z_k)/T)} \tag{6}$$

其中，z_o和$z_{o'}$分别代表同一原始样本不同增强版本的特征表示. T 是温度参数，此外$1_{[k\neq o]}$是指示函数，当$k \neq o$时，取值为 1，否则为 0. 这确保了在求和时不会考虑与自身的相似度.

接着，采用对比学习方法对失真样本进行学习，在失真样本之间使用对比学习，可以让模型更好的理解失真数据的特征. 同样也是希望将同一个失真样本的两个不同增强版本的特征相互拉近，不同失真样本的特征拉远. 因此，对于失真样本，该文定义了如下的损失函数:

$$l_{\text{con\_dis}} = -\ln \frac{\exp(\text{sim}(z_d, z_{d'})/T)}{\sum_{k=1}^{2N} 1_{[k\neq d]} \exp(\text{sim}(z_d, z_k)/T)} \tag{7}$$

其中，z_d和$z_{d'}$分别代表同一失真样本不同增强版本的特征表示. $1_{[k\neq d]}$代表指示函数，当$k \neq d$时，取值为 1，否则为 0.

通过最小化上述的两个损失函数，可以促使模型更加准确地识别并区分原始样本与失真样本中的特征. 这样的优化策略不仅增强了模型在原始数据上的表现，而且提高了其对失真数据的稳健性.

随后，为了进一步加强模型对原始和失真样本的区分能力，该文还添加了一个新的损失函数，该损失函数的目标是最小化原始样本和其对应的失真样本之间的余弦相似度，拉远它们在特征空间中的距离. 具体而言，原始样本和失真样本之间的新损失$l_{distance}$定义如下:

$$l_{distance} = \frac{z_o \cdot z_d}{\|z_o\| \|z_d\|} \tag{8}$$

z_o和z_d分别表示经过编码后的原始样本及其对应失真样本的特征向量. 这一操作不仅增强了模型的判别能力，同时确保了模型能够清晰地区分原始样本和失真样本. 该文自监督预训练的最终任务就是最小化整体目标函数L_{total}.

$$L_{total} = l_{dis} + l_{pace} + \alpha \cdot l_{con\_org} + \beta \cdot l_{con\_dis} + \gamma \cdot l_{distance} \tag{9}$$

使用失真类型预测、对比学习和播放速度预测作为辅助任务为模型提供了丰富和多样性的学习目标，这种多任务学习策略允许模型从不同角度捕获数据的内在结构，从而更全面地理解其背后的模式. 预测失真类型这一辅助任务指导模型学习到不同类型失真的特征表示，对比学习使模型充分学习了样本之间的相似性和差异性，预测播放速度这一辅助任务使模型能够捕捉视频中的动态变化和时序特征. 这些优势有助于提高模型的性能，使其更加适应实际应用中的挑战.

2 基于多尺度时空融合的视频质量评估模型

该文提出的 VQA 模型如图 2 所示. 模型首先通过多尺度时空特征提取模块提取视频的时空特征，随后引入 SimAM 注意力机制[21]来提高模型对关键特征的关注度. 接着，使用 Swin Transformer 的编码器部分对视频帧内的空间特征进行更深层次的建模. 最后，通过线性回归模型，将提取的高级特征映射为视频质量评分.

在该文提出的视频质量评估模型中，首先设计了一个多尺度时空特征提取模块，该模块首先利用 3D-CNN 对视频输入进行处理，以有效地捕获视频的时空特征F_{ts}. 3D-CNN 通过在时间维度上进行卷积操作，能够同时处理视频帧的空间信息和时间序列信息，从

而帮助模型更好地理解视频的动态变化和时序关系. 在此基础上, 采用基于 ResNet50[24]骨干网络的多尺度架构, 进一步提取时空特征. 具体而言, 从 ResNet50 骨干网络的最后四层提取特征后, 这些特征通过 1 × 1 卷积操作进行处理, 以统一不同特征层级的通道数, 随后通过上采样操作将低分辨率的深层特征图尺寸增大, 并与高分辨率的浅层特征图逐元素相加进行融合. 经过上采样后, 各层特征通过自适应平均池化操作映射到相同的尺度, 最后, 将这些经过自适应池化的各层特征进行拼接, 从而得到最终的多尺度时空特征表示. 该多尺度时空特征提取模块通过结合 3D-CNN 和多尺度架构, 不仅能够提取视频的动态时序特征, 还能在不同空间尺度上捕捉细节信息, 为模型提供了丰富的时空特征表示. 此外, 模型引入了 SimAM 注意力机制[21]以有选择性的关注视频中的关键信息, 进一步增强特征的代表性和区分性. 经过多尺度时空特征提取模块和 SimAM 注意力机制的处理后, 得到了增强后的特征F_{en}, 该特征充分考虑了视频中各个尺度和关键帧的信息.

此外, 为了进一步提高空间特征编码的效率和准确性, 该文在模型中集成了 Swin Transformer. Swin Transformer 通过其分层和窗口化的架构, 不仅提高了处理视频帧空间特征的效率, 还增强了模型对帧内复杂结构的识别能力. 这种架构特别适合处理大规模视频数据, 它在提高计算效率的同时, 能够全面捕捉视频的时空特征. 随后将增强后的特征F_{en}输入到 Swin Transformer 编码器中, 以提取出深层次的特征F_{final}. 通过多尺度时空特征提取模块、注意力机制等模块的结合使用, 最终得到的特征F_{final}能够更全面地理解视频的时空特性, 从而提高视频质量评估的整体准确性.

在整个特征提取模块得到了最终的特征F_{final}后, 需要将这些特征映射到具体的质量分数. 为了实现这一映射, 该文选择了线性回归模型, 这是因为它的简单性和对于此任务的有效性. 该文使用一个单层全连接网络作为回归模型, 其具有单个输出神经元, 用于直接预测视频的质量分数. 视频质量分数计算如下:

$$Q_{pred} = FC(F_{final}) \tag{10}$$

在模型微调阶段, 首先加载预训练好的模型权重, 然后使用公共数据集进行微调, 以使模型能够更好地适应具体任务和数据集. 微调过程中, 采用可微分的 PLCC 损失函数[26]来调整模型参数, 以提高模型在预测视频质量方面的精确性和一致性, 从而在最终的测试中获得更优的结果. 微调的损失函数如公式(11)所示:

$$loss\ l = \frac{(1-PLCC(Q_{pred},Q_{gt}))}{2} \tag{11}$$

其中, Q_{pred}和Q_{gt}分别代表预测的质量分数和对应的主观质量分数.

3　实验结果与分析

3.1 测试数据集与评价指标

为了全面评估该文方法在处理真实失真和合成失真上的能力, 选择了四个关键的VQA数据集进行测试. 在真实失真类别中, 选取了 KoNViD-1k 和 LIVQ-VQC 这两个数据集. KoNViD-1k 由 1200 个从多种设备和分辨率中捕获的用户生成内容视频组成, 而 LIVQ-VQC 由 80 名移动相机用户拍摄的 585 个视频场景构成. 在合成失真类别中, 选用了 LIVE 和 CSIQ 这两个数据集. LIVE 数据库, 由德克萨斯州奥斯汀分校提供, 包含 160 个视频, 覆盖了四种主要的失真类型. 与此同时, CSIQ 数据库则包括了六种不同失真类型的 228 个视频. 这四个数据库的选择旨在确保模型能够在真实和合成失真条件下进行全面验证和评估, 数据集的详细内容如表 1 所示.

表 1　所用数据集详细参数说明

数据集	类型	视频个数	分辨率	时长
LIVE	合成失真数据集	160	768x432	10 秒
CSIQ	合成失真数据集	228	832x480	10 秒
KoNVid-1k	真实失真数据集	1200	540p	8 秒
LIVE-VQC	真实失真数据集	585	240p-1080p	10 秒

为了确保模型性能的准确评估, 该文选取了两种关键的评价标准：皮尔逊线性相关系数(Pearson Linear Correlation Coefficient, PLCC)和斯皮尔曼秩相关系数(Spearman Rank Order Correlation Coefficient,, SROCC). PLCC 考量的是预测值与真实值之间的线性关系, 其公式表示为：

$$PLCC = \frac{\sum_i (S_i-\mu_S)(S_i'-\mu_S')}{\sqrt{\sum_i (S_i-\mu_S)^2 \sum_i (S_i'-\mu_S')^2}} \tag{12}$$

其中, S_i和S_i'分别代表第 i 个视频的真实质量评分和预测评分, 而μ_S和μ_S'是真实评分和预测评分的均值. 而

表 2 在合成失真数据集上的性能比较（最好结果加粗显示, 次优结果加下划线显示）

		LIVE 数据集		CSIQ 数据集	
		SROCC	PLCC	SROCC	PLCC
有监督方法	RIRNet	0.7516	0.6877	0.7957	0.7715
	VIDEVAL	0.6716	0.6739	0.6498	0.6592
	SIONR	0.5977	0.6585	0.6567	0.7009
	Q-Boost	——	——	——	——
自监督方法	CSPT	0.7276	0.8196	0.7398	0.7093
	CONVIQT	0.6220	0.5950	0.7660	0.7490
	SelfVQA	0.8699	0.8850	0.9069	0.8949
	Ours	**0.8952**	**0.9087**	**0.9105**	**0.9075**

表 3 不同 VQA 方法针对单一失真类型的 SROCC 值对比（最好结果加粗显示, 次优结果加下划线显示）

	LIVE 数据集				CSIQ 数据集					
	WL	IP	H264	MPEG2	H264	WLPL	MJPEG	SNOW	AWGN	HEVC
TLVQM	0.6071	0.4857	0.7167	0.8571	0.8667	0.6333	0.8929	0.8286	0.8	0.75
VIDEVAL	0.5394	0.6429	0.7	0.85	0.8061	0.3019	0.8818	0.8214	0.7857	0.8
SIONR	0.6777	0.76	0.8714	0.7952	0.7943	0.6914	0.4857	0.5249	0.4572	0.6457
RIRNet	0.5809	0.76	0.519	0.8986	0.8742	0.6685	0.8971	0.8742	0.9086	0.8628
CSPT	0.7619	0.8285	0.8304	0.7763	0.8736	0.7143	0.7191	0.8857	0.8743	0.8476
CONVIQT	0.595	0.486	0.738	0.81	0.817	0.533	0.8	0.867	0.8	0.717
SelfVQA	0.7074	0.8051	0.8285	0.9095	0.92	0.897	**0.9429**	**0.9428**	0.8857	0.9543
Ours	**0.8146**	**0.8429**	**0.9048**	**0.9424**	**0.9581**	**0.9152**	0.9321	0.9381	**0.9111**	**0.9717**

SROCC 则对预测值的单调趋势进行描述, 其公式表示为:

$$SROCC = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)} \tag{13}$$

其中, d_i是第 i 个视频真实评分和预测评分之间的排名差距, 而 n 是视频的总数. 理论上, 如果 SROCC 和 PLCC 的值接近于 1, 那么这表示模型具有出色的预测能力. 在进行 PLCC 的计算前, 该文按照[25]中的方法, 利用四参数逻辑函数对客观评分进行非线性映射, 使其与主观评分对应.

3.2 实验设置

该文的实验在配备有两块 GeForce RTX 3090 显卡的服务器上进行, 网络是在 PyTorch 框架内实现. 在自监督预训练过程中, 采用了随机梯度下降优化器, 并设置了初始学习率为 0.001、动量为 0.9, 权重衰减为 0.005. 为了调整学习率, 该文使用了步长学习率调度器, 每 6 个训练周期后, 学习率会乘以一个衰减因子 0.1. 在微调阶段, 该文使用的初始学习率为 0.0003, 训练周期数为 100. 损失函数权重系数α、β、γ的取值分别为 1、1 和 0.5. 温度参数 T 设定为 0.07. 该文将公共数据集百分之八十用于微调, 百分之二十用于测试, 为了保证结果的鲁棒性, 这种微调测试过程重复五次, 然后使用评估指标的平均值, 作为最终的评估结果.

3.3 性能比较

为了全面评估该文方法在各种场景中的表现, 选择了一系列流行的质量评估模型作为基准进行比较, 包括：SIONR[5]、RIRNet[6]、BVQI[9]、Q-Boost[14]、VIDEVAL[15]. 考虑到近年来自监督学习方法在许多领域都取得了令人瞩目的成果, 该文特意挑选了一些采

表 4　在真实失真数据集上的性能比较（最好结果加粗显示, 次优结果加下划线显示）

		KoNVid-1k 数据集		LIVE-VQC 数据集	
		SROCC	PLCC	SROCC	PLCC
有监督方法	RIRNet	0.7475	0.7388	0.7056	0.7108
	VIDEVAL	0.7830	0.7790	0.7416	0.7493
	SIONR	0.82	0.8146	0.7613	0.7769
	Q-Boost	0.8010	0.8030	0.7410	0.7930
	BVQI	0.7600	0.7600	0.7840	0.7940
自监督方法	CSPT	0.8145	0.8104	0.7604	0.7628
	VISION	0.5980	0.5970	0.6760	0.7010
	CONVIQT	**0.8510**	**0.8490**	0.8080	0.8170
	SelfVQA	0.8179	0.8160	0.7647	0.7894
	Ours	0.8389	0.8489	**0.8133**	**0.8251**

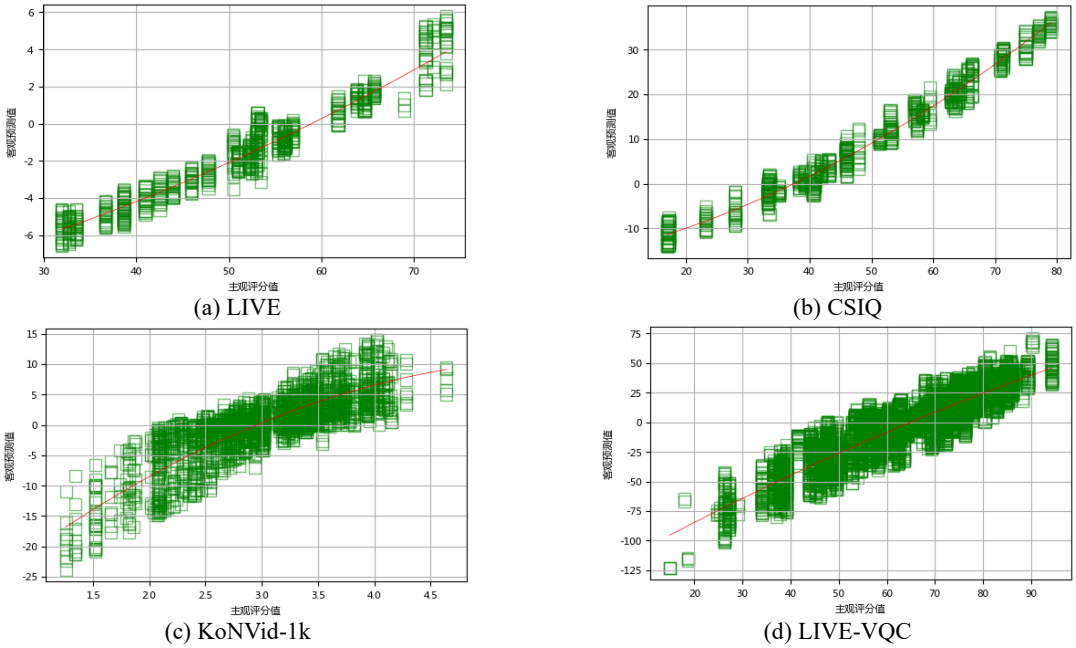
用自监督策略的模型, 包括：CSPT[17]、VISION[18]、CONVIQT[19] 、SelfVQA[20]确保了比较的公正性和全面性, 这不仅有助于深入了解自监督方法在此类任务中的优势, 还能证明该文方法在与其他先进技术进行比较时具有竞争力.

从表 2 可以看出, 在合成失真数据库 LIVE 和 CSIQ 上, 该文提出的模型在 SROCC 和 PLCC 两个指标上均达到了最高的分数, 显示出了卓越的性能. 与目前最好的自监督学习方法 SelfVQA 相比, 在 SROCC 上分别提高了 2.91%和 0.39%, 在 PLCC 上分别提高了2.68%和1.41%. 这得益于将失真类型和对比学习作为辅助任务, 因为合成失真数据集通常包含固定的失真类型, 如模糊、噪声、压缩等. 当模型在训练时使用失真类型作为一个辅助任务, 它可以更有效地学习到这些特定失真的特征, 从而在测试阶段具有更好的鉴别和预测能力, 对比学习也可以帮助模型提高对视频质量的敏感性.

为了深入评估该文提出的模型在处理不同失真类型时的性能, 又进一步对各个单一失真类型进行了单独训练和测试. 将该文方法与其他 VQA 方法在 LIVE 和 CSIQ 数据集上进行了对比, 实验结果如表 3 所示. 可以观察到与所有其他 VQA 方法相比, 该文方法在 LIVE 数据集上对所有类型的失真都显示了显著的性能提升, 并在 CSIQ 数据集的大部分失真类型上都取得了最好的结果. 这些结果验证了该文使用失真类型预测和对比学习作为自监督辅助任务的有效性. 然而, 在 MJPEG 和 SNOW 失真上取得了次优的结果, 与预训练数据集对这两种失真类型的样本覆盖不足有关.

从表 4 可以看出, 在真实失真数据集 LIVE-VQC 上, 该文方法领先于其他所有方法, 与自监督学习方法 SelfVQA 相比, 在 SROCC 上提高了 6.3%, 在 PLCC 上提高了 4.52%, 这得益于该文在预训练阶段采用的预测播放速度方法能够很好的捕获视频的动态特征, 这些特征在移动相机拍摄的视频中尤为重要. 但是在 KoNViD-1k 数据集上, 结果低于 CONVIQT, 这种性能差异的原因在于该文使用的预训练数据远少于 CONVIQT, CONVIQT 使用了 6 万个视频进行预训练, 大量的数据量提高了其模型对不同内容的适应性和鲁棒性. 此外, KoNViD-1k 数据集包含多种不同的内容, 这些内容与模型的训练数据有很大的差异, 进而就会导致模型在此数据集上的泛化能力下降.

图 3 展示了模型在 LIVE、CSIQ、KoNVid-1k 和 LIVE-VQC 数据集上的可视化表现. 这些散点图直观地展示了模型预测的质量分数与各数据集的主观质量评分之间的关系. 为了更加清晰地展示这种关系, 该文使用了一条拟合曲线以描绘客观预测值和主观评分值之间的趋势. 从散点图中可以看出, 预测值围绕着拟合曲线分布, 这表明该文所提方法在不同数据集上都能够与主观评分保持较好的一致性, 进一步说明了该文所提方法性能的准确性和可靠性.



(a) LIVE (b) CSIQ

(c) KoNVid-1k (d) LIVE-VQC

图 3 不同数据集的主观质量分数与所提模型预测值的散点图

3.4 跨数据集实验

理想的视频质量评估模型应当具备良好的泛化性, 能够在面对从未接触过的数据样本时, 依旧保持稳定的评估效能. 为了验证该研究所提模型的这一能力, 该文采取了跨数据集的测试方法, 具体而言从一个数据集上进行训练, 然后在另外三个数据集上进行测试, 以检验模型在未经训练的数据集上的表现.

表 5 跨数据集实验结果

训练	测试	SelfVQA	ours
LIVE	CSIQ	0.3315	0.3159
	KoNVid-1k	0.3088	**0.3835**
	LIVE-VQC	0.4077	**0.5397**
CSIQ	LIVE	0.4049	**0.4104**
	KoNVid-1k	0.6243	**0.6316**
	LIVE-VQC	0.578	**0.5851**
KoNVid-1k	LIVE	0.1504	**0.1591**
	CSIQ	0.4896	0.4468
	LIVE-VQC	0.6584	**0.6735**
LIVE-VQC	LIVE	0.2165	**0.2419**
	CSIQ	0.3733	0.3702
	KoNVid-1k	0.7052	**0.7088**
平均结果		**0.4374**	**0.4555**

跨数据集的实验结果如表 5 所示. 总体而言, 我们的模型展现了卓越的竞争力. 例如, 当模型在 KoNVid-1k 上训练并在 LIVE-VQC 上测试时, 该模型性能相比于基线模型提升了 2.29%. 然而, 由于 CSIQ 数据集包含 6 种复杂多样的失真类型, 当模型在其他数据集（如 LIVE 或 KoNVid-1k）上训练并在 CSIQ 数据集上测试时, 可能未充分学习到这些复杂失真的特征, 导致性能不佳. 尽管在某些特定测试条件下表现有所下降, 但是整体来看, 该模型在所有测试条件下的平均性能提升了 4.14%. 这一结果表明该模型能够将从一个数据集中学习到的特征, 有效迁移到其他数据集. 跨数据集的实验结果不仅突显了自监督预训练对于增强模型鲁棒性的重要性, 也验证了该模型捕捉到的特征具有较强的泛化能力.

3.5 消融实验

3.5.1 自监督辅助任务消融实验

为了验证该文提出的自监督学习方法中各辅助任务的有效性, 探究失真类型、播放速度以及对比学习作为辅助任务对模型性能的影响, 本节将展示一系列消融实验结果. 具体而言, 按照辅助任务的不同组合进行了六组实验, 并在 LIVE 数据集上进行测试, 以观察模型在不同配置下的性能变化.

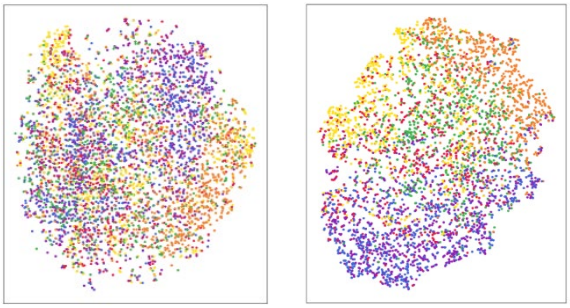
通过表 6 的结果可以观察到, 当单独移除播放速度预测任务时(对比第五行和第七行), SROCC 下降了 4.59%, PLCC 下降了 6.16%. 这一发现证明了播放速度

预测在捕捉视频的动态变化和时序特征方面发挥了关键作用，对提高模型的整体性能至关重要. 其次，当单独去除失真类型任务时(对比第六行和第七行)，模型的 SROCC 下降了 3.90%, PLCC 下降了 4.20%，这说明失真类型预测同样对模型的评估能力有显著影响. 而当去除对比学习任务时(对比第四行和第七行)，尽管模型性能的下降幅度相对较小，但它在辅助失真类型的特征区分上起着积极作用，这一点从表中第一行和第五行的比较中得以显现.

表 6　自监督辅助任务消融实验结果

失真类型预测	播放速度预测	对比学习结合余弦相似度	SROCC	PLCC
√			0.8405	0.8376
	√		0.8584	0.8488
		√	0.8276	0.8272
√	√		0.8809	0.8786
√		√	0.8541	0.8527
	√	√	0.8602	0.8705
√	√	√	0.8952	0.9087

同时该文使用 t-SNE 图更加直观的展示了对比学习对特征分布的影响，可视化结果如图 4 所示，不同颜色的点代表不同的失真类型. 图 4(a)展示了未使用对比学习时的特征空间分布，而图 4(b)则显示了对比学习引入后，不同失真类型的特征在一定程度上被模型区分并聚集. 这一结果不仅体现了对比学习促使模型更加准确地区分不同失真样本特征，也说明了其对于整体模型性能提升的贡献.



(a)未加入对比学习　　(b)加入对比学习

图 4　t-SNE 可视化，不同颜色代表不同的失真类型

从表 6 中也可以发现，仅使用单一辅助任务的效果不如完整的多任务框架（对比第一行和第七行、第二行和第七行、第三行和第七行），但仅使用播放速度预测任务的性能表现优于仅使用其他两种辅助任务的情形（对比第一行、第二行和第三行）. 上述分析进一步证实了播放速度预测在提升视频质量评估准确性方面的重要性.

上述消融实验结果不仅证实了所提出的各辅助任务对于提升模型性能的作用，也验证了该文所用的多任务自监督学习框架的有效性.

3.5.2 多尺度时空特征提取模块消融实验

为了深入理解该文所提多尺度时空特征提取模块在 VQA 模型中的作用，该研究对所提出的 VQA 模型中的多尺度时空特征提取模块进行了单独的消融分析.

表 7　多尺度时空特征提取模块消融实验结果

多尺度时空特征提取模块	SROCC	PLCC
×	0.8359	0.8578
√	0.8952	0.9087

实验结果如表 7 中所示，使用多尺度时空特征提取模块的模型在性能上与不使用多尺度时空特征提取模块相比，前者在 SROCC 上比后者提高了 7.09%，在 PLCC 上提高了 5.93%. 这一显著的性能提升证明了多尺度时空特征提取模块引入了更丰富的、更多层次的语义信息，增强了 VQA 模型的时空建模能力，从而提高了模型在视频质量评估方面的有效性.

4　结论

该文针对视频质量评估任务中的整体时空特征捕捉不足的问题，提出了一个综合利用播放速度预测、失真类型识别和对比学习的自监督学习框架. 该框架通过引入多个辅助任务以预训练 VQA 模型，旨在更准确地捕获视频的动态变化和时序特征. 此外，该文采用了多尺度时空特征提取模块、SimAM 注意力机制和 Swin Transformer 以增强时空建模能力. 为了对提出的方法进行验证，该文在四个数据集上进行了广泛的实验，实验结果证明了该文方法相比现有的自监督方法具有显著的性能优势，特别是在处理复杂动态场景时. 这些结果不仅证明了自监督学习在视频质量评估任务中的有效性，也显示了该方法在捕捉动态视频质量关键特征方面的潜力. 在接下来的工作中，计划探索更多种类的自监督辅助任务，以提高模型对于各种失真

类型和视频内容的适应性, 同时优化模型架构, 以进一步提高 VQA 任务的准确性和效率.

参考文献

1 Wang Z, Rehman A. Begin with the end in mind: A unified end-to-end quality-of-experience monitoring, optimization and management framework[C]//SMPTE 2017 Annual Technical Conference and Exhibition. SMPTE, 2017: 1-11.

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

3 Xu M, Chen J, Wang H, et al. C3DVQA: Full-reference video quality assessment with 3d convolutional neural network[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 4447-4451.

4 Soundararajan R , Bovik A C . Video Quality Assessment by Reduced Reference Spatio-Temporal Entropic Differencing[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2013, 23(4):684-694.

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

6 Chen P, Li L, Ma L, et al. RIRNet: Recurrent-in-recurrent network for video quality assessment[C]//Proceedings of the 28th ACM International Conference on Multimedia. 2020: 834-842.

7 Huang D J, Kao Y T, Chuang T H, et al. SB-VQA: A Stack-Based Video Quality Assessment Framework for Video Enhancement[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 1613-1622.

8 You J, Lin Y. Efficient Transformer with Locally Shared Attention for Video Quality Assessment[C]//2022 IEEE International Conference on Image Processing (ICIP). IEEE, 2022: 356-360.

9 Wu H, Liao L, Wang A, et al. Towards robust text-prompted semantic criterion for in-the-wild video quality assessment[J]. arXiv preprint arXiv:2304.14672, 2023.

10 Lin L, Wang Z, He J, et al. Deep Quality Assessment of Compressed Videos: A Subjective and Objective Study[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022.

11 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. 2023: 1066-1076.

12 施文娟, 孙彦景, 左海维, 等. 基于视频自然统计特性的无参考移动终端视频质量评价[J]. 电子与信息学报, 2018, 40(1): 143-150.

13 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.

14 Zhang Z, Wu H, Ji Z, et al. Q-Boost: On Visual Quality Assessment Ability of Low-level Multi-Modality Foundation Models[J]. arXiv preprint arXiv:2312.15300, 2023.

15 Tu Z, Wang Y, 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.

16 Wang J, Jiao J, Liu Y H. Self-supervised video representation learning by pace prediction[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16. Springer International Publishing, 2020: 504-521.

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

18 Mitra S, Soundararajan R. Multiview Contrastive Learning for Completely Blind Video Quality Assessment of User Generated Content[C]//Proceedings of the 30th ACM International Confer-ence on Multimedia. 2022: 1914-1924.

19 Madhusudana P C, Birkbeck N, Wang Y, et al. Conviqt: Contras-tive video quality estimator[J]. IEEE Transactions on Image Pro-cessing, 2023.

20 Jiang S, Sang Q, Hu Z, et al. Self-Supervised Representation Learning for Video Quality Assessment[J]. IEEE Transactions on Broadcasting, 2022.

21 Yang L, Zhang R Y, Li L, et al. Simam: A simple, parameter-free attention module for convolutional neural networks[C]//International conference on machine learning. PMLR, 2021: 11863-11874.

22 Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.

23 Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding[J]. arXiv preprint arXiv:1807.03748, 2018.

24 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

25 Jochen Antkowiak, T Jamal Baina, France Vittorio Baroncini, Noel Chateau, France FranceTelecom, Antonio Claudio França Pessoa, F Stephanie Colonnese, Italy Laura Contin, Jorge Ca-viedes, and France Philips. 2000. Final report from the video quality experts group on the validation of objective models of video quality assessment march 2000. (2000).

26 Wu H, Chen C, Hou J, et al. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling[C]//European Con-ference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 538-554.

27 P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, “Image quality assessment using contrastive learning,” IEEE Trans. Image Process., vol. 31, pp. 4149–4161,2022.