高动态范围还是标准动态范围？缩放和压缩视频的主观与客观研究

Joshua P. Ebenezer，*IEEE*学生会员，Zaixi Shang，*IEEE*学生会员，Yixu Chen，Yongjun Wu，Hai Wei，Sriram Sethuraman，Alan C. Bovik，*IEEE*会士

arXiv:2304.13162v1 [eess.IV] 25 Apr 2023

摘要——**我们开展了一项大规模研究，考察人们对经过缩放并采用不同压缩等级的高动态范围（HDR）和标准动态范围（SDR）视频的感知质量评价，这些视频在三种不同的显示设备上播放。与标准动态范围视频相比，高动态范围视频能够呈现更宽的色域、更好的对比度、更亮的白色和更暗的黑色。尽管按照传统预期，高动态范围视频的质量优于标准动态范围视频，但我们发现，观众对高动态范围和标准动态范围视频的偏好在很大程度上取决于显示设备，以及分辨率缩放和比特率。为研究这一问题，我们收集了67名志愿者对356段视频的23,000多条质量评分；这些志愿者使用OLED、QLED和LCD电视观看视频。鉴于有必要衡量这些场景下的视频质量，例如为缩放、压缩以及标准动态范围与高动态范围之间的选择提供依据，我们在新建的数据库上测试了若干知名的全参考和无参考视频质量模型。为推动这一问题的研究进展，我们还开发了一种名为HDRPatchMAX的新型无参考模型，它结合经典失真统计量和对位深度敏感的失真统计量，准确度高于现有指标。**

索引词—**高动态范围、视频质量评价、视频压缩**

I. 引言

High动态范围（HDR）视频采用更深的位深度来表示更高和更低的亮度，并且比标准动态范围（SDR）视频具有更宽的色域。然而，要充分发挥HDR的优势，显示设备必须具备准确呈现高对比度以及动态范围两端极值的技术。SDR视频采用BT709 [1]中所述的幂律进行伽马编码。在电视上显示时，这些值会被解码，并使用查找表根据电视的显示能力进行调整。尽管SDR标准最初是为最大显示亮度为100尼特的阴极射线电视机制定的，但现代显示设备可以利用其整个亮度范围（通常远高于100尼特）来显示SDR内容，因为SDR内容的数字值是相对的，而不是场景相关的。SDR数字值与电视能够显示的内容之间的确切映射因产品而异，通常属于专有技术。

另一方面，遵循HDR10标准的视频具有场景相关性和绝对性。PQ电光传递函数（用于HDR10标准）规定了显示器必须针对特定数字值显示的绝对亮度值。如果所需显示的绝对亮度值高于电视机能够显示的最高亮度值，则会在HDR内容上应用一种称为电-电传输函数（EETF）的色调映射函数，以避免高光出现剪切，同时确保亮度值在峰值处平滑滚降。不同电视机所采用的EETF各不相同，通常也是专有的。

由于标准动态范围和高动态范围信号在支持HDR的显示器上的显示方式存在差异，某一内容的SDR版本的平均亮度可能高于HDR版本，具体取决于两者的调色和显示方式。例如，HDR中最大亮度为200尼特的内容，可能会以这样的方式进行SDR调色，使最大亮度的数字值达到255。当HDR版本和SDR版本在峰值亮度为1000尼特的支持HDR的设备上显示时，SDR版本呈现的峰值亮度可能远高于200尼特，而HDR版本则会以200尼特的峰值亮度显示。因此，SDR版本可能看起来更亮，但也可能出现褪色或过饱和现象。更高的峰值亮度或更高的平均亮度并不是HDR内容比SDR更具吸引力的唯一原因。事实上，标准动态范围视频可能会出现饱和度异常、色带、低对比度等缺陷，而这些缺陷在高动态范围视频中不太容易出现。

除这些差异外，高动态范围视频使用10位表示（存储于16位中），而标准动态范围视频采用8位表示。高动态范围视频所需的字节数是相同内容的标准动态范围视频的两倍。由于这一点，高动态范围视频可能更容易受到压缩伪影的影响。

压缩、对比度表示、颜色表示和亮度之间的权衡，使得高动态范围视频和标准动态范围视频质量的感知评估取决于内容和显示设备。为了更好地理解这些权衡，我们对具有相同内容的高动态范围视频和标准动态范围视频进行了详细的主观和客观评估。在主观研究中，我们招募了67名参与者，让他们观看并评价25种源内容中的356个高动态范围视频和标准动态范围视频的质量；这些视频由x265编码器采用不同的缩放和压缩组合进行处理。我们还在新的主观数据库上评估了客观的全参考（FR）和无参考（NR）视频质量评估（VQA）。此外，我们还介绍了一种新的无参考视频质量评估模型的设计，用于预测高动态范围视频和标准动态范围视频的质量。

J.P. Ebenezer、Z. Shang 和 A.C. Bovik 任职于图像与视频工程实验室（德克萨斯大学奥斯汀分校），地址为 Austin,TX, 78712, USA，电子邮箱：joshuaebenezer@utexas.edu。Y. Chen、Y. Wu、H. Wei 和 S. Sethuraman 任职于 Amazon Prime Video。J.P. Ebenezer 和 Z. Shang 对本项工作贡献相同

II. 相关工作

据我们所知，目前尚无研究在动态范围、分辨率和压缩等级均发生变化的情况下比较视频的主观质量。现有数据库，如LIVE Livestream [2],LIVE ETRI [3],LIVE YTHFR [4], AVT UHD [5], 和APV LBMFR [6]，研究了专业制作的标准动态范围视频在下采样、压缩和源失真条件下的主观质量。其他数据集，包括Konvid-1k [7], YouTube UGC [8], 和LSVQ [9]，研究了标准动态范围用户生成内容的质量。由于对显示设备、分辨率和比特率的要求较为宽松，用户生成内容数据库通常远大于研究专业级内容质量的数据库，因为前者可以通过众包在线开展。LIVE HDR [10],LIVE AQ HDR [11], 和APV HDR Sports [12]是近期建立的数据库，研究了经过不同分辨率和比特率下采样及压缩的专业制作高动态范围视频的质量。

上述各数据库研究的都是标准动态范围或高动态范围视频的质量，而非两者兼有，且这些视频都经过了失真处理。在此，我们展示了首项主观研究，比较了相同内容的高动态范围视频与标准动态范围视频在经过下缩放和压缩处理后的质量。我们在多种采用不同技术、性能各异的显示设备上开展了这项研究。

像我们所开展的这类研究所获得的人类主观评分，被视为视频质量的黄金标准，但开展此类研究成本高昂，且难以扩展。不过，客观视频质量指标旨在通过设计和训练，实现视频质量的自动预测，因此成本较低且易于扩展。这些指标分为两类：全参考（FR）和无参考（NR）模型。全参考视频质量评价模型需要同时输入原始视频和失真视频，以评估失真视频的质量。无参考指标在预测质量时只能获取失真视频，因此设计这类指标更具挑战性。无参考视频质量评价模型适用于视频源检查，也适用于没有可用源视频时的质量测量。

峰值信噪比衡量参考帧与同一帧的失真版本之间的峰值信号噪声比。结构相似性 [13] 结合亮度、对比度和结构特征来预测失真图像的质量。视频多方法评估融合 [14] 对视频帧小波系数的统计特性以及失真造成的细节损失进行建模。SpEED [15]衡量参考视频与失真视频的带通系数熵差异。STRRED [16]对受失真影响的时空视频小波系数统计特性进行建模。STGREED [17] 衡量失真产生的时间熵和空间熵差异，以对具有不同帧率和比特率的视频质量进行建模。

BRISQUE [18], VBLIINDS [19], VIDEVAL [20],RAPIQUE [21], ChipQA [22], HDR ChipQA [23] 和 NIQE [24] 是依赖视觉感知神经统计模型的无参考视频质量指标。众所周知，无失真视频经过处理后会遵循某些规律性的统计特征使用视觉神经模型。失真会以可预测的方式改变经过感知处理的视频统计特征，从而可以设计出准确的视频质量评估模型。RAPIQUE 将基于这些模型开发的特征，与预训练深度网络提供的（语义）视频特征相结合。TLVQM [25]使用各种空间和时间滤波器及启发式方法，对压缩、模糊和闪烁等常见失真进行显式建模。

III. 主观研究详情

该研究以向67名受试者展示356个视频的方式进行。这些视频由25个独特且未经失真的内容生成，这些内容还被转换为标准动态范围，且全部经过下缩放和HEVC压缩的组合处理。研究采用了三种不同的电视技术。

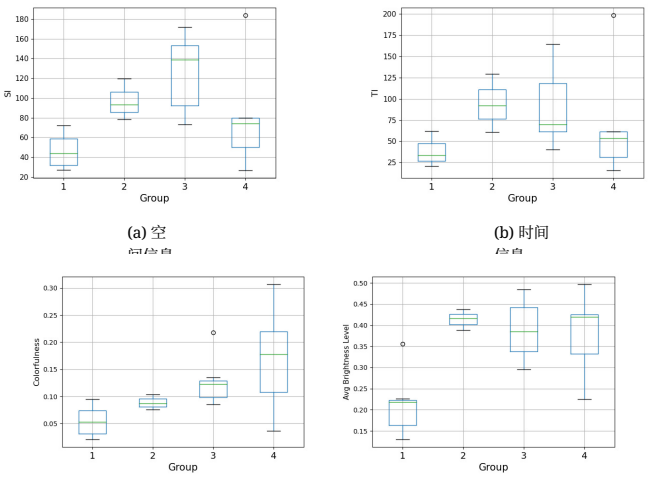
A. 源序列

这 25 个源序列可分为 4 组：视频点播（VoD）、体育直播、高动态范围演示视频和 SJTU 视频。其中 3 个源序列是取自既有高动态范围视频质量评估数据库（LIVE HDR、LIVE AQ HDR 和 APV HDR Sports）的“锚点”序列，用于校准这些数据库的数据并将其与当前数据库的数据合并。所有视频均采用 BT2020 [26] 色域，并使用 SMPTE ST2084 [27] 光电传递函数（OETF）进行量化，该函数也称为感知量化器（PQ）。所有源序列的时长均在 7 秒至 10 秒之间，静态元数据符合 HDR10 标准。以下将详细介绍各视频类别：

1) 视频点播（*VoD*）：这 7 个视频由专业人员为视频点播流媒体服务拍摄并调色。这些视频的高动态范围和标准动态范围版本由 Amazon Studios 制作。其中一个类别内容是来自 LIVE AQ HDR数据库的“锚点”视频。标准动态范围版本由专业调色师以创作意图为指导制作并手动调色。

2）体育直播：这4段现场直播体育视频由广播机构在举办足球和网球比赛的体育场内进行专业拍摄。该类别中的一段内容是来自APV HDR Sports数据库的“锚点”视频。由于现场直播的低延迟要求，这些视频在HDR和SDR格式下均使用预设查找表（LUT）进行调色。HDR调色使用的LUT是亚马逊专有LUT，而用于将HDR转换为SDR的LUT则是开源NBC LUT。所有视频最初均为YUV422 10位格式，之后转换为受限的YUV420 10位格式。

*3) HDR*演示视频：这些视频是从4kmedia.org收集的8个开源视频。视频由电视制造商制作并调色，用于展示高动态范围相较于标准动态范围的优势，因此这些内容具有较高的对比度和色彩丰富度，以吸引观众的注意。源视频采用有限范围的YUV420 10位格式。随后使用NBC查找表将其转换为标准动态范围版本。



(a) 空 (b) 时间

(c) 色彩丰富度 (d) 平均亮度

图1：内容描述符，其中 1=点播视频、 2=现场直播体育视频、 3=HDR演示视频和 4=SJTU视频。详情请见正文。

*4) SJTU*视频：该类别包含六个源内容。这些内容取自开源的SJTU HDR视频序列数据集。它们使用Sony F65摄像机拍摄，并使用S-Gamut查找表进行调色。视频采用有限范围的YUV420 10位格式，并同样使用NBC查找表转换为标准动态范围版本。该类别中的所有源内容也存在于LIVE HDR数据库中，但其失真版本有所不同。

B. 内容描述符

针对每个视频序列计算了空间信息（SI）、时间信息（TI）、色彩丰富度和平均亮度水平，并按所属组别绘制于图1中。与其他类别相比，VoD内容的平均亮度水平、SI、TI和色彩丰富度指数均较低。Live内容具有较高的TI。Demo Videos具有较高的对比度、色彩丰富度和亮度，这是因为它们旨在展示HDR的能力。SJTU视频包含许多不同场景，因此具有很高的变化性。其SI和TI低于第2组和第3组，不过其中的一段烟花视频在该组的四项内容描述符上均属于异常值。

C. 源序列的处理

VoD内容包括演播室调色HDR和SDR源版本，而其他内容类别则是原始HDR版本，这些版本还使用NBC查找表转换为SDR版本。HDR和SDR版本采用x265编码器的限比特率恒定码率因子（CRF）方法进行编码。每种分辨率对应的最大比特率（最大比特率）、CRF和缓冲区大小列于表I中。

九个EPL6内容的高动态范围视频（来自现场直播体育类别）取自APV HDR Sports数据库，九个NightTraffic内容的高动态范围视频（来自SJTU类别）取自LIVE HDR数据库，九个BTFB-01h05m40s内容的高动态范围视频取自LIVE AQ HDR数据库。每组九个视频均包括原始版本和八个压缩版本。这27个“锚点”视频用于估计从先前高动态范围数据库中的评分到新HDR与SDR数据库中评分的映射关系。每个锚点内容还包含6个压缩SDR版本，其最大比特率和分辨率如表I所示，另有一个原始SDR版本。因此，每个锚点内容分别对应16个视频序列（9个高动态范围版本和7个SDR版本）。

其余22个源内容各自对应一个无损高动态范围版本、一个无损标准动态范围版本、六个压缩高动态范围版本和六个压缩SDR版本，共计308个视频。这些视频与锚定内容中的视频合计，在新数据库中构成356个视频。因此，数据库中共有181个高动态范围视频和175个标准动态范围视频。

D. 显示设备

本研究选取了三台电视机：65英寸 Samsung S95T（TV1）、65英寸 Samsung Q90（TV2）和55英寸 Amazon Fire TV（TV3）。所有电视机均能够接收并显示HDR10内容。峰值亮度是指电视机显示HDR时能够达到的最大亮度。不同电视机宣称具有不同的“峰值亮度”能力，但它们可能只能在短时间内使屏幕的一小部分达到所宣称的数值。这样做是为了防止显示设备受损并降低功耗，这一过程称为自动亮度限制。我们将峰值亮度定义为覆盖屏幕2%区域的白色矩形显示时的瞬时亮度，该数值由 [28] 报告。

Samsung S95 配备量子点有机发光二极管（QD-OLED）显示屏。每个像素都能自行发光，因此该设备可以呈现极高的对比度。其 BT.2020 色彩空间覆盖率为 86.93%，峰值亮度为 1028 cd/m^2。Samsung Q90T 配备量子点显示屏和垂直取向（VA）LED 背光。量子点能够精准地发出红光和绿光，因此比标准 LED 电视呈现更加鲜艳的色彩。Samsung Q90 的 BT.2020 色彩空间覆盖率为 67.24%，峰值亮度为 1170 cd/m^2。Amazon Fire TV 是一款入门级 VA LED 电视，Rec 2020 色域覆盖率为 54.25%，峰值亮度为 230cd/m^2。Samsung Q90T 和 Amazon Fire TV 均具备全阵列局部调光（FALD）特征，可调暗屏幕中需要呈现较暗效果区域的背光亮度，以提升对比度。不过，由于采用背光，较亮区域的光线仍可能“渗漏”到较暗区域，从而影响对比度。另一方面，Samsung S95 配备 OLED 显示屏，因此可以单独控制每个像素，获得优于 FALD 所能实现的对比度。不过，由于配备 LED 背光和能够增强光线的量子点，Samsung Q90T 的峰值亮度可以高于 S95。

表I

分辨率	HDR的CRF	SDR的CRF	最大比特率（千比特每秒）	缓冲区大小（千字节）
540p	25	30	800	2000
720p	17	24.5	3000	6000
1080p	20	27	3000	8000
2160p	25	28	6000	15000
1440p	15	19	10000	20000
2160p	17.5	22	15000	22500

使用搭载 NVIDIA 3090 图形处理器、运行 Windows 10 操作系统的 Windows 个人电脑，通过 HDMI 2.1 电缆驱动电视机。个人电脑和电视机均启用了高动态范围。屏幕分辨率设为 3840x2160，刷新率设为 30 Hz。使用 VLC 媒体播放器播放视频。

E. 受试者

共有67名德克萨斯大学奥斯汀分校的学生自愿参与这项人体研究。所有受试者的年龄均在20至28岁之间。约三分之二的受试者为男性，其余三分之一自我认定为女性。人口统计调查显示，73%的受试者自我认定为亚裔，20%自我认定为白人，7%自我认定为黑人。

其中，22名受试者被安排观看TV1，21名被安排观看TV2，24名被安排观看TV3。为消除偏差，没有受试者被告知其他电视或研究的性质。所有受试者在佩戴矫正镜片（如有需要）的情况下，均通过了石原色盲测试和斯内伦视力测试。

F. 主观测试设计

我们按照国际电信联盟 ITU-R BT 500 中所述的方法，采用单刺激法开展了这项研究。13 [29]每个视频向每名受试者播放一次，并在视频播放完毕后立即收集受试者给出的质量评分。视频以随机顺序播放，且未标明参考视频和失真视频，也未对二者进行区别处理。为减少记忆偏差，不允许相同内容的视频在观看顺序中相邻出现。每个质量评分均通过一个不可见的1至100整数刻度收集，该刻度配有一个连续滑块，并设有5个文字标记：“差”、“较差”、“一般”、“良好”和“优秀”。受试者使用鼠标操作滑块。研究分为两场进行，每场约40分钟，两场之间至少间隔24小时，以减少观看者疲劳。

在每次测试环节开始前，都会进行培训，使受试者通过观看不属于该数据库的视频熟悉实验设置。向每位受试者展示六个内容相同但压缩程度不同的视频，其中三个为高动态范围视频，三个为标准动态范围视频，以较为充分地体现数据库中的质量范围。研究人员向受试者演示了如何使用评分机制。在培训期间，研究人员说明了应如何根据对视频质量的主观评估进行评分，同时避免对美学内容作出判断。为避免对参与者产生偏差，未提供有关该研究的其他说明或细节。

IV. 主观分析

A. 内部相关性

三个组的内部相关性按如下方式计算。设被试 i 对电视 TV t 上视频 j 给出的评分为 u_{ijt}。Z 评分的计算方式为

$$Z_{ijt} = \frac{u_{ijt} - \frac{\Sigma_j u_{ijt}}{N}}{\sqrt{\frac{\Sigma_j (u_{ijt} - \Sigma_j u_{ijt})^2}{N}}}. \tag{1}$$

将被试随机分成两个相同规模的组，并计算每个组中所有被试对每个视频给出的平均 Z 评分。针对不同的随机分组进行 100 次试验，计算这两个组对所有视频给出的评分之间的相关性。TV1 观众的被试间相关性中位数为 0.95，TV2 为 0.94，TV3 为 0.93。这些数据表明内部一致性和可靠性很高。

B. 平均意见分的计算

平均意见分（MOS）采用国际电信联盟910 [30]提出的最大似然估计方法计算。MOS 被建模为一个随机变量

$$U_{ijn} = \Psi_{jn} + \Delta_i + \nu_i X \tag{2}$$

其中，Ψ_{jn} 表示在电视n上观看视频 j 时的真实质量，Δ_i 表示受试者 i的偏差，ν_i 表示受试者 i的不一致性，而 $X \sim N$，1)为独立同分布的高斯随机变量。给定评分 u_{ijn}，通过将 Ψ_{jn}，Δ_i，ν_i 视为自由参数并求解，使模型（2）与观测到的平均意见分最佳拟合，从而估计每台电视上每个视频的真实评分。具体而言，Ψ_{jt} 通过使用牛顿-拉夫森求解器最大化观测值的对数似然来确定。

通过计算失真视频与参考视频之间的平均意见分差异，得到差分平均意见分（DMOS），具体如下：

$$DMOS_{jt} = \Psi_{jt} - \Psi_{j_0 t}, \tag{3}$$

其中，j_0 是与在电视TVn上观看的视频 j 相对应的参考视频索引。因此，平均意见分和差分平均意见分均针对每个视频分别在每台电视上计算。

C. 分数分析

图2、图3、图4和图5展示了4种内容的20个视频（每个源组包含1种内容，每种内容包含5个视频）的平均意见分与比特率之间的关系。Forge视频来自第1组，EPL来自第2组，Color DJ来自第3组，Porsche来自第4组。第一列展示了源内容的标准动态范围版本的屏幕截图。每条线上的点依次对应于540p、720p、1080p、1440p、2160p和2160p，顺序递增，与表I一致。

Forge视频展示了铁匠铺中锻造一把剑的场景。烧红的铁发出的光线与周围及背景中的黑暗和阴影形成了强烈对比。在TV1上，Forge视频的高动态范围版本评分高于标准动态范围版本，因为高动态范围版本中的细节更加清晰。然而，在TV2和TV3上，由于它们显示对比度的能力较弱，高动态范围视频更高的压缩程度抵消了对比度提升所带来的相对质量增益，因此标准动态范围版本的评分更高。由于高动态范围与标准动态范围采用了不同的显示方式（如引言中所述），以及视频调色方式的差异，高动态范围版本看起来比标准动态范围版本更暗、更昏沉。较低的平均亮度也可能导致观众感知到的高动态范围视频内容质量较低。高动态范围版本较低的平均亮度、TV2和TV3降低的对比度，以及场景较高的空间复杂度，也可能解释了为什么在较高比特率下，标准动态范围版本的评分仍然更高。对于从第2组获取的EPL视频，也可以观察到类似现象：在TV1上，高动态范围版本的评分更高，而在TV2和TV3上，较低比特率下标准动态范围版本的评分更高。在较高比特率下，高动态范围版本的评分更高，因为尽管足球场景具有较高的时间复杂度，压缩伪影的视觉影响较小。

ColorDJ 视频的标准动态范围版本存在颜色和亮度范围较广的问题，因此出现了过饱和和过曝。另一方面，高动态范围版本由于位深度更高、色域更广，没有出现过饱和或过曝。这或许可以解释为什么高动态范围版本在所有电视机上的总体评分都更高。

Porsche 视频中的汽车有鲜艳的红色车漆，在高动态范围版本中得到了准确呈现，但在标准动态范围版本中看起来过于饱和。该视频在空间和时间上的复杂度也不高，这可能是因为高动态范围在呈现对比度和鲜艳色彩方面的能力更强，其效果超过了压缩带来的影响。不过，该视频的 720p 高动态范围版本在 1515 kbps 下编码，其平均意见分与比特率曲线出现了急剧下降。编码器决定以 1300 kbps 对 1080p 版本进行编码，但 1080p 版本的评分仍高于 720p 版本。这很可能是因为该内容的空间复杂度较低，缩放伪影比压缩伪影更为明显。

图6绘制了所有视频的高动态范围版本与标准动态范围版本的平均意见分之间的平均差异与每台电视的最大比特率之间的关系。如图所示，在较低比特率下，标准动态范围的评分高于高动态范围；而在较高比特率下，差异变为正值。1080p内容在最大比特率为3000 kbps时以及2160p内容在最大比特率为6000 kbps时曲线出现下降，这表明，为了获得最佳比特率阶梯，这些分辨率应采用更高的比特率进行编码。需要注意的是，本研究纳入这些压缩等级是为了代表广泛的质量范围，而不是代表最佳比特率阶梯。对于TV1，即使在较低比特率（3000 kbps）下，高动态范围的表现也优于标准动态范围，并且在最高比特率下，质量差异增大到5个平均意见分单位。对于TV2和TV3，在较低比特率下，标准动态范围的质量优于高动态范围，但随着比特率提高，差异逐渐减小，并在较高比特率下变为略微正值。这也表明，显示设备的性能会显著影响高动态范围内容相对于标准动态范围内容的质量。

D. 数据库合并

使用来自 LIVE HDR、LIVE AQ HDR 和 APV HDR Sports VQA数据集的 27 个锚定视频的评分，将这些数据库中的评分映射到当前数据库。对于每台电视机，均使用逻辑函数，将这些数据库中锚定视频的评分映射为同一视频在 LIVE HDRvsSDR 数据库中的评分。逻辑函数为

$$f(x) = \frac{a-b}{(1+\exp\frac{-(x-c)}{s})} + b, \tag{4}$$

其中， x 表示先前数据库中视频的评分， $f(x)$表示针对特定电视机将评分映射到 LIVE HDRvsSDR 数据库中的映射函数， a、b、c和 s分别使用各数据库中的锚定视频求解。由于每台电视机和每个（先前）数据库都用于生成不同的拟合函数，因此根据数据共得到九个函数。将 LIVE HDR数据库和 LIVE AQ HDR数据库中的评分映射到 LIVE HDRvsSDR 数据库所使用的三台电视机所得评分的函数，分别绘于补充材料中的图 12 和图 13 中。由于专有性原因，无法展示 APV HDR Sports数据库的映射和结果。

根据锚定视频的评分推导出这些函数，可以将三个 LIVE 数据库合并为一个包含 1066 个视频的大型数据库，这些视频均采集于受控实验室环境中。

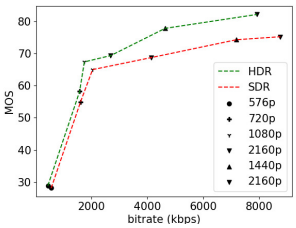
V. 客观评估

A. 全参考视频质量评估

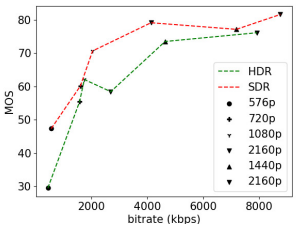
我们在新建的数据库上测试了峰值信噪比（PSNR）、结构相似性指数（SSIM） [13], 多尺度结构相似性指数（MS-SSIM） [31],视频多方法评估融合（VMAF） [14], 时空约简参考熵差（STRRED） [16],空间高效熵差（SpEED） [15], 以及时空广义熵差（STGREED） [17]等视频质量模型。我们通过计算算法预测分数与真实差分平均意见分之间的斯皮尔曼秩相关系数（SRCC），评估了这些全参考视频质量评估算法。



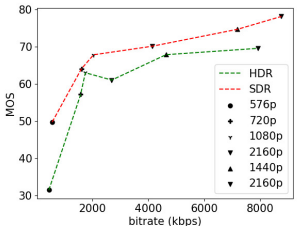
(a) 屏幕截图 from 视频“Forge”。



（b）TV1的平均意见分与比特率



（c）TV2的平均意见分与比特率

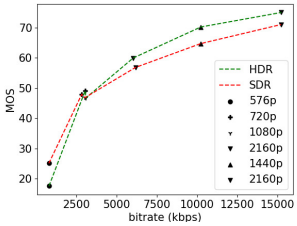


（d）TV3 的平均意见分与比特率

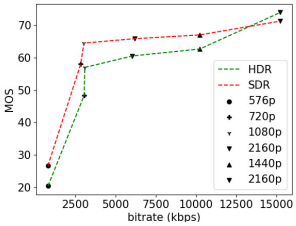
图 2：使用“Forge”视频对三台受测电视机进行的平均意见分与比特率关系图。



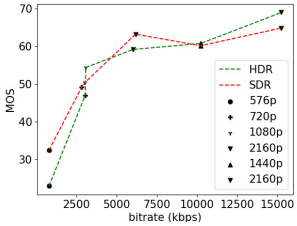
（a）视频“EPL”的屏幕截图。



（b）TV1 的平均意见分与比特率

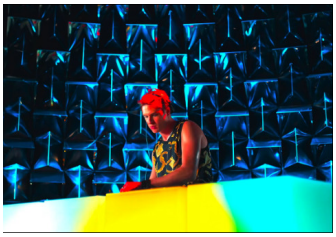


（c）TV2 的平均意见分与比特率

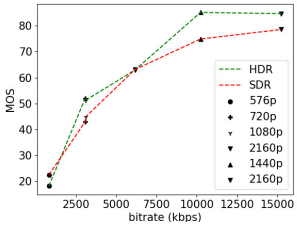


（d）TV3 的平均意见分与比特率

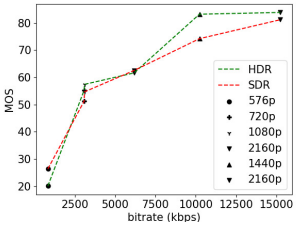
图3：“EPL”视频在三台受测电视上的平均意见分与比特率关系图。



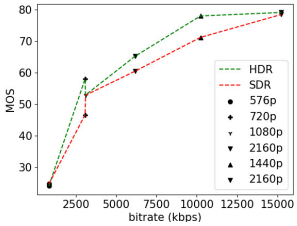
（a）视频“Col- orDJ”的屏幕截图。



（b）TV1 的平均意见分与比特率关系



（c）TV2 的平均意见分与比特率关系



（d）TV3 的平均意见分与比特率关系

图4：“ColorDJ”视频在三台受测电视上的平均意见分与比特率关系图。

我们还使用逻辑函数将预测分数拟合到差分平均意见分（DMOS）。

$$l(s) = \frac{\beta_1 - \beta_2}{1 + \exp(-\frac{(x-\beta_3)}{\beta_4}) + \beta_5}, \tag{5}$$

然后按照标准做法，根据拟合分数与差分平均意见分之间的皮尔逊线性相关系数（LCC）和均方根误差（RMSE）进行计算 [32]。结果如表II所示。

可以看出，全参考指标在该任务上的表现并不理想，其中视频多方法评估融合（VMAF）在TV1、TV2和TV3上的斯皮尔曼秩相关系数（SRCC）最高，分别为0.56、0.55和0.59。此外，由于这些模型均未对显示设备进行建模，因此它们在三台电视机上的预测结果完全相同，只有STGREED例外，因为STGREED需要针对每台电视分别训练一个支持向量回归（SVR）模型。这些结果凸显了针对高动态范围（HDR）和标准动态范围（SDR）内容开展全参考视频质量评估研究的必要性。图7中绘制了预测结果与TV1所得分数之间的关系。

B. HDRPatchMAX

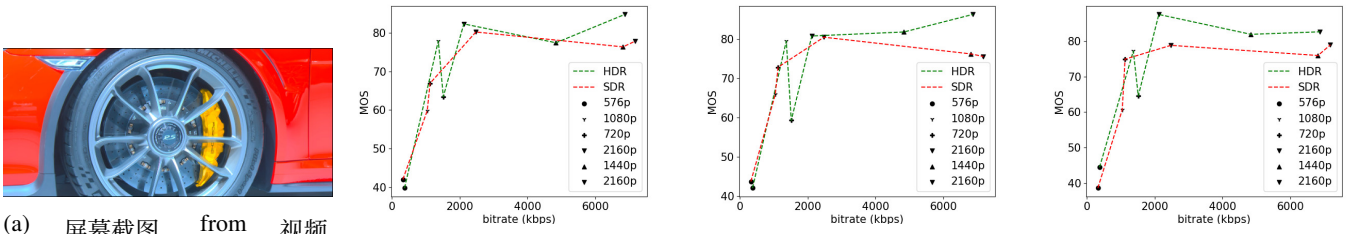
受现有模型性能不佳的启发，我们还设计了一种新的高动态范围无参考视频质量评估算法，称为 HDR-PatchMAX。HDRPatchMAX 利用了与标准动态范围和高动态范围质量以及运动感知相关的特征。

*1) NIQE*特征：HDRPatchMAX中的第一组特征与NIQE中的特征相同 [24]。NIQE可以单独作为无参考指标使用，但我们对每个视频所有帧中的36个特征和距离度量取平均值，并使用所得的37个特征进行后续训练。

2) PatchMAX： NIQE 得到的 37 个特征在 LIVE HDR、LIVE AQ HDR 和 LIVE HDRvsSDR 数据集上表现出色，这促使我们开发了一种基于图像块的特征提取方法。我们首先将帧划分为大小为 $P \times P$ 的不重叠图像块，并计算其均值减除对比度归一化ˆ（MSCN）系数。视频帧或图像块的亮度通道的 MSCN 系数 $V[i,j,k_0]$ $V[i,j,k_0]$ 定义为：

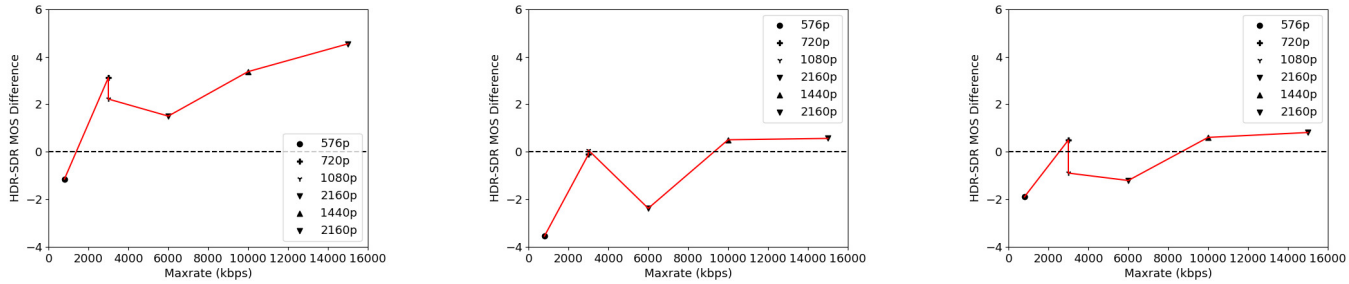
$$\hat{V}[i,j,k_0] = \frac{V[i,j,k_0] - \mu[i,j,k_0]}{\sigma[i,j,k_0] + C}, \tag{6}$$

其中 $i \in 1,2..,M$, $j = 1,2..,N$ 为空间索引， M和N 分别为图像块的高度和宽度，常数



(a) 屏幕截图 from 视频 “Porsche”。 (b) TV1 (c) TV2 (d) TV3

图5：“Porsche”视频在三台受测电视上的平均意见分与比特率关系图。



(a) TV1 的平均意见分与比特率关系图 (b) TV2 的平均意见分与比特率关系图 (c) TV3 的平均意见分与比特率关系图

图6：不同电视上对应的高动态范围视频与标准动态范围视频的平均意见分之平均差异与最大比特率的关系图。

表II：全参考视频质量评估算法在新的 LIVE HDRvsSDR 数据库上的 SRCC、LCC 和均方根误差中位数。括号中为标准差。每种电视的最佳结果以粗体显示。

方法	TV1			TV2			TV3		
	斯皮尔曼秩相关系数↑	皮尔逊线性相关系数↑	均方根误差↓	斯皮尔曼秩相关系数↑	LCC↑	均方根误差↓	斯皮尔曼秩相关系数↑	LCC↑	均方根误差↓
PSNR	0.1056	0.3170	18.8984	0.0380	0.2673	19.1349	0.1351	0.2918	17.7954
STRRED [16]	0.1865	0.3228	18.8593	0.1287	0.2691	19.1252	0.1166	0.2331	18.0922
SpEED [15]	0.3209	0.4535	17.7103	0.2326	0.3796	17.4674	0.1771	0.3095	16.6531
多尺度结构相似性 [31]	0.4172	0.4391	17.9224	0.3071	0.3502	17.3550	0.3181	0.3544	17.3256
结构相似性 [13]	0.4807	0.4921	17.3459	0.4184	0.4439	17.7940	0.3818	0.4021	17.0344
视频多方法评估融合 [14]	**0.5654**	**0.5850**	**16.1606**	**0.5547**	**0.5810**	**16.1618**	**0.5971**	**0.6057**	**14.8045**
STGREED [17]	0.4873(0.0908)	0.5114(0.0643)	17.3710(1.0645)	0.4719(0.0902)	0.4814(0.0635)	17.3219(1.1705)	0.4471(0.1191)	0.4663(0.0758)	16.1962(1.2141)

C 赋予数值稳定性，并且在

$$\mu[i,j,k_0] = \sum_{m=-L}^{m=L}\sum_{l=-L}^{l=L} w[m,l]V[i+m,j+l,k_0] \tag{7}$$

and

$$\sigma[i,j,k_0] = (\sum_{m=-L}^{m=L}\sum_{l=-L}^{l=L} w[m,l](V[i+m,j+l,k_0] - \mu[i,j,k_0])^2)^{\frac{1}{2}} \tag{8}$$

分别是亮度的局部加权空间均值和标准差。权重 $w = \{w[m,l]|m = -L、\ldots\ldots、L、l = -L、\ldots\ldots、L\}$ 是一个二维圆对称高斯加权函数，该函数采样至 3 个标准差处，并重新缩放至单位体积； $K = L = 3$。

MSCN 运算过程中计算出的标准差会在每个图像块上取平均值，并用作各图像块对比度的替代指标。随后，根据各图像块的标准差相对于同一帧中其他图像块标准差的百分位数，将图像块分为三组。高对比度图像块定义为标准差高于 $(100-T)^{\text{th}}$ 百分位数的图像块；中对比度图像块定义为标准差介于第 T^{th} 百分位数和第 $(100-T)^{\text{th}}$ 百分位数之间的图像块；低对比度图像块则是标准差低于第 T^{th} 百分位数的图像块。

这些图像块的MSCN系数可以可靠地建模为广义高斯分布（GGD），定义为

$$g_1(x;\alpha;\beta) = \frac{\alpha}{2\beta\Gamma(\frac{1}{\alpha})}\exp[-(\frac{|x|}{\beta})^{\alpha}] \tag{9}$$

其中 Γ（。）是伽马函数：

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1}\exp(-t)dt. \tag{10}$$

对MSCN系数进行最佳拟合得到的形状参数和方差参数 α 和 β 将被提取并用作质量感知特征。

随后，计算每个图像块中相邻像素对的乘积，以捕捉它们之间的相关性

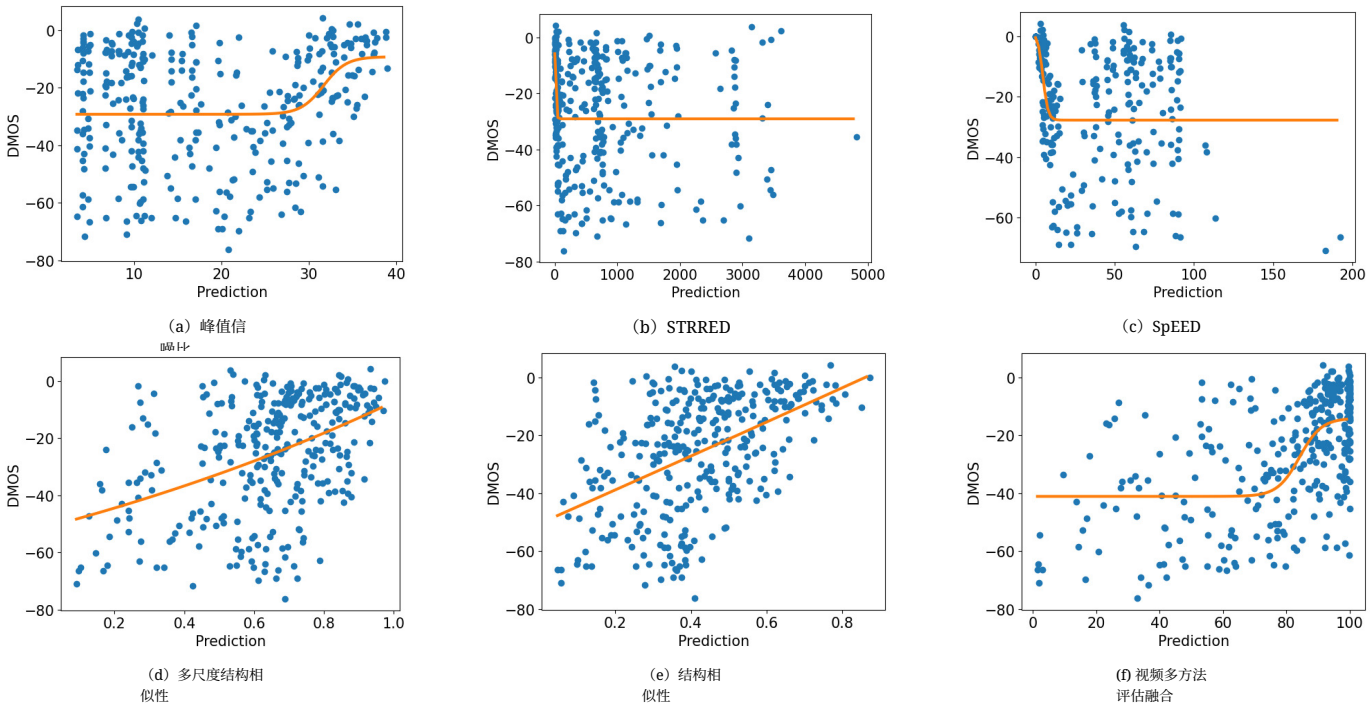


图 7：TV1 中所示视频的差分平均意见分与全参考视频质量评估预测值的散点图，橙色曲线为逻辑拟合。

如下:

$$\begin{aligned} H[i,j,k_0] &= \hat{V}[i,j,k_0]\hat{V}[i,j+1,k_0] \\ V[i,j,k_0] &= \hat{V}[i,j,k_0]\hat{V}[i+1,j,k_0] \\ D_1[i,j,k_0] &= \hat{V}[i,j,k_0]\hat{V}[i+1,j+1,k_0] \\ D_2[i,j,k_0] &= \hat{V}[i,j,k_0]\hat{V}[i+1,j-1,k_0]. \end{aligned} \tag{11}$$

将这些数据建模为服从非对称广义高斯分布（AGGD）：

$$g_2(x;\nu,\sigma_l^2,\sigma_r^2) = \begin{cases} \frac{\nu}{(\beta_l+\beta_r)\Gamma(\frac{1}{\nu})}\exp(-(-\frac{x}{\beta_l})^\nu) & x<0 \\ \frac{\nu}{(\beta_l+\beta_r)\Gamma(\frac{1}{\nu})}\exp(-(\frac{x}{\beta_r})^\nu) & x>0, \end{cases} \tag{12}$$

其中

$$\beta_l = \sigma_l\sqrt{\frac{\Gamma(\frac{1}{\nu})}{\Gamma(\frac{3}{\nu})}} \quad \text{and} \quad \beta_r = \sigma_r\sqrt{\frac{\Gamma(\frac{1}{\nu})}{\Gamma(\frac{3}{\nu})}}, \tag{13}$$

其中，ν 控制分布形状，σ_l和 σ_r 控制众数两侧的离散程度。参数（η，ν，σ_l^2，σ_r^2）由对式（11）中每一组成对乘积的直方图进行最佳非对称广义高斯分布拟合后提取，其中

$$\eta = (\beta_r-\beta_l)\frac{\Gamma(\frac{2}{\nu})}{\Gamma(\frac{1}{\nu})}. \tag{14}$$

将每个图像块的最佳拟合参数提取为质量特征，并分别对三类图像块的参数取平均，从而得到总计54个特征。该过程在两个尺度上进行，因此总计得到108个特征。除此之外，我们还计算这些特征在每个不重叠的五个连续帧分组上的平均时间标准差，并使用这额外的108个特征作为时空质量特征。图像块大小 P 和百分位阈值 T 被视为超参数，并根据表III所示的结果，将其分别选为 $P=20$ 和 $T=10$。

采用对比度分离特征聚合有三个重要原因。首先，高动态范围与标准动态范围的主要差异之一在于对局部对比度的呈现方式。受限于 8 位深度，标准动态范围帧对边缘和细节的呈现能力不如高动态范围帧。然而，与对比度可见性提升相对应的特征响应，可能会被帧中其他区域的特征响应所掩盖，而这些区域未必能从高动态范围更精确的量化中受益。按对比度明确区分特征响应，可以避免这种掩盖效应。例如，在图 8(b) 中，高动态范围帧的高对比度区域以白色显示，中等对比度区域以灰色显示，低对比度区域以黑色显示。图 8(e) 则以相同方式显示同一帧的标准动态范围版本。在标准动态范围版本中，女性 T 恤和前景男子西装背部的区域曝光不足；我们提出的基于对比度的分割将其标示为低对比度区域。在高动态范围版本中，这些区域曝光正常，低对比度区域则仅位于背景的单调墙面上。因此，这些区域的统计特征也会有所不同，并与质量相关。

其次，在存在对比度极低的区域（例如天空）时，NSS建模会受到影响。大多数自然场景都充分验证了关于无失真帧MSCN系数高斯性的既有观察结果，但完全缺乏纹理或细节的非常平滑区域可能是这一自然规律的罕见例外。将此类区域与对比度和纹理更高的区域

进行区分，可能会提高这些模型的有效性和效力。

第三，对比度掩蔽是一种重要的视觉现象，即在存在高对比度纹理时，失真可能变得不那么明显，这为分别分析具有不同对比度的区域提供了进一步的动机。MSCN 运算用于建模人类视觉系统早期阶段发生的对比度增益掩蔽，而明确区分具有不同对比度的区域，则有助于进一步建模这一效应。

图9绘制了压缩HDR视频中某一帧的、按对比度分组的不同图像块的MSCN系数，以及相应的SDR帧。系数明显不同，因此对这些分布进行的最佳广义高斯分布拟合也将相应地具备质量感知能力。

3) HDRMAX：HDRMAX特征集 [33]的提取过程如下：首先，将每一帧的亮度值通过一种扩展非线性进行处理，该方法最初由HDR ChipQA引入。首先，利用每个窗口中的最小值和最大值，将大小为 20×20 的重叠窗口中的值线性重缩放至 $[-1, 1]$ 。随后，将这些值输入定义如下的扩展非线性：

$$f(x;\delta) = \begin{cases} \exp(\delta x) - 1 & x >= 0 \\ 1 - \exp((-\delta x)) & x < 0 \end{cases} \tag{15}$$

其中 $\delta = 4$ 基于先前在 [33] LIVE HDR数据库上的实验。该非线性通过抑制中间范围的值并增强亮度范围两端的极值来放大局部对比度。由于 HDR 擅长表现对比度，这一操作会加剧固有的压缩失真，并迫使后续特征响应聚焦于这些失真。

例如，在图 8(c) 和图 8(f) 中，HDRMAX 非线性分别应用于图 8(a) 和图 8(d) 所示的帧。所得图像中的局部对比度在整个帧范围内都得到增强。标准动态范围图像中存在欠曝光的区域（如女性的 T 恤）以及过曝的区域（如后方房间中的日光灯），其亮度值接近恒定，因此会被拉伸至极端，并形成凸显其缺陷的清晰边界。HDR 版本中的相应区域则更加平滑，而这些区域中的失真凸显了 HDR 表现对比度的增强能力。

对非线性处理后的帧计算MSCN系数，并将其建模为广义高斯分布；同时，将其相邻乘积建模为非对称广义高斯分布，从而得到18个特征。这些特征在两个尺度上提取。此外，我们还计算了每五帧组成的每个不重叠帧组中这些特征的平均标准差，从而得到72个特征。

4) 时空芯片：我们使用时空（ST）梯度芯片对时空信息进行建模，该方法首次在[34] 中提出，并在 [22]中得到进一步发展。梯度包含有关边缘和失真的重要信息，而ST芯片是梯度视频的时空切片，旨在以质量感知的方式捕获时间信息。

表III：改变图像块大小和百分位数阈值对TV1的SRCC中位数的影响。

图像块大小 P	百分位阈值 T	SRCC
20	10	0.8347(0.0828)
20	20	0.8244(0.0902)
20	30	0.7917(0.1144)
120	10	0.8035(0.0997)
120	20	0.8003(0.1061)
120	30	0.8055(0.1035)
240	10	0.7838(0.1197)
240	20	0.7853(0.1123)
240	30	0.7838(0.1197)

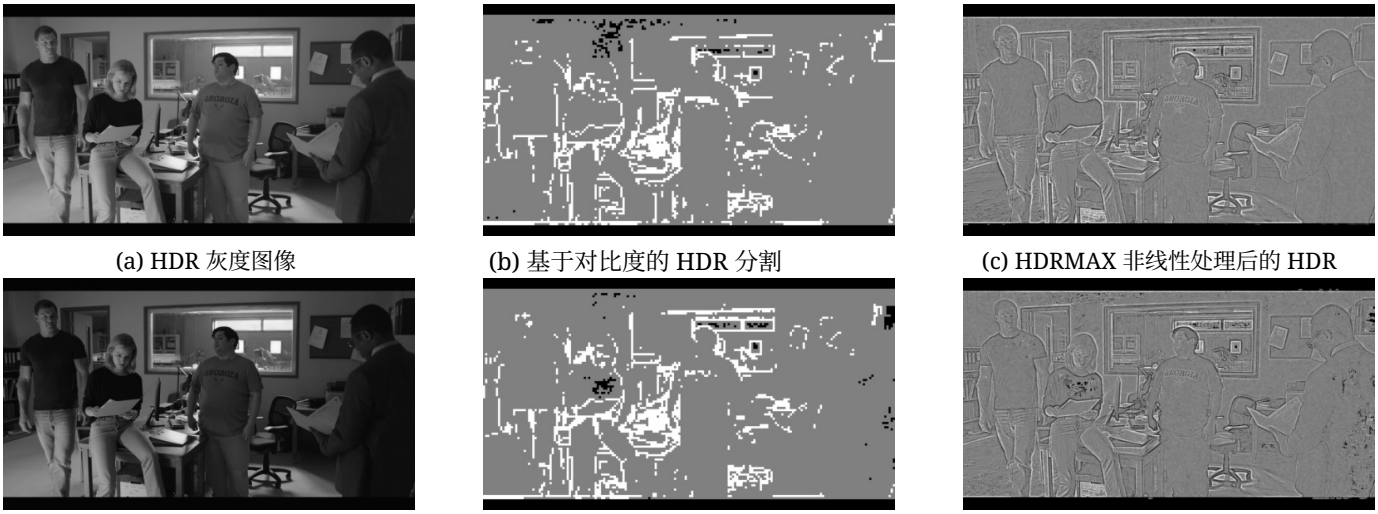
首先使用3×3 Sobel算子计算每一帧的梯度幅值。根据公式（6）计算梯度幅值的MSCN系数。随后应用时间带通滤波器 k（t）：

$$h(t) = t(1 - at)\exp(-2at)u(t). \tag{16}$$

该简单线性滤波器模拟了外侧膝状体（LGN） [35]中发生的时间去相关过程。视觉信号从LGN传递到V1区，其中神经元对局部区域内的运动敏感。受此启发，通过对六个方向进行网格搜索，并选择峰度最接近高斯分布的切片，从每个空间位置选取梯度视频MSCN系数的时空切片。

具体而言，令ˆ $G[i, j, k]$ 表示时间索引为 k的视频帧梯度幅值的MSCN系数。对于 $k = T$，定义一个以时间 T结束的5帧块，记为 $G_T = G[i, j, k]$；$k = T - 4, \ldots T$。然后，在每个ˆ时空体 G_T中定义六个5x5时空切片或芯片，它们与全部5帧相交，彼此间隔 $\frac{\pi}{6}$ 角，并受到以下约束：每个ˆ芯片的中心与体 G_T的中心重合，且每个芯片的法向量位于空间平面内。垂直于运动方向的时空芯片能够捕捉沿该向量方向的物体运动，这里我们默认小型时空体的运动是平移的。我们此前已表明，高质量视频中垂直于局部运动方向的时空芯片的自然带通统计量可靠地服从高斯分布，而指向其他方向的芯片则偏离高斯性 [22]。这一观察源于无失真帧的MSCN系数所具有的高斯性。因此，在每个 $5 \times 5 \times 5$ 体中，我们选择超额峰度最小的 5×5 时空芯片。ST Gradient芯片系数建模为GGD，其相邻成对乘积建模为AGGD，由此得到18个特征。在两个尺度上重复该过程，共得到36个特征。

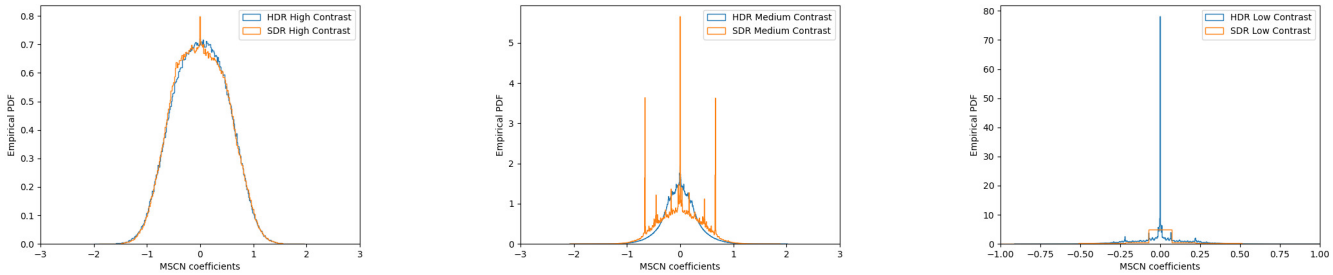
HDRPatchMAX因此由37个NIQE特征、108个PatchMAX特征、72个HDRMAX特征和36个STGradient Chip特征组成。



(a) HDR 灰度图像 (b) 基于对比度的 HDR 分割 (c) HDRMAX 非线性处理后的 HDR

(d) SDR 灰度图像 (e) 基于对比度的 SDR 分割 (f) 经过 HDRMAX 非线性处理后的标准动态范围。

图 8：HDR 与 SDR 比较。



（a）高对比度图像块 （b）中对比度图像块 （c）低对比度图像块

图 9：比较不同对比度组的 HDR 与 SDR MSCN系数。

C. 无参考视频质量评估

我们在新的LIVE HDRvsSDR 数据库上评估了RAPIQUE [21], 、BRISQUE [18],、TLVQ [25],、VBLIINDS [19], 、ChipQA [22],和HDR ChipQA [23],，以及HDRPatchMAX。同样，我们报告了SRCC、PLCC和均方根误差指标。算法采用随机森林回归器进行训练。我们还测试了支持向量回归器，但由于随机森林回归器的性能更好，最终选择了随机森林回归器。视频按照80:20的比例划分为训练集和测试集，以确保相同内容的视频出现在同一数据集中。在训练集上进行交叉验证，以选择随机森林的最佳超参数（估计器数量和特征数量）。随后，使用最佳超参数的随机森林在训练集上拟合，并在测试集上进行评估。该过程使用不同的随机训练测试划分重复100次，并在表IV中报告中位数结果及标准差。HDRPatchMAX在所有电视机上均优于其他无参考视频质量评估算法。参数 P 和 T 根据PatchMAX特征集的性能进行选择，因为 P 和 T发生了变化，具体结果见表III。

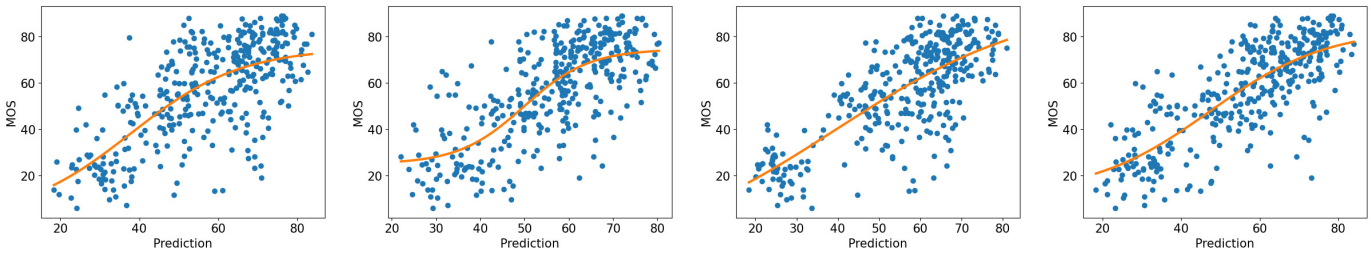
无参考视频质量评估算法预测结果的散点图与TV1上所显示视频获得的平均意见分进行对比，如图10所示。散点图的绘制方法是：针对测试集中的每个视频，在100次训练测试划分中，将平均意见分与无参考视频质量评估算法生成的平均质量预测值进行绘图，并以橙色显示公式5中的逻辑拟合结果。与其他算法相比，HDRPatchMAX的预测结果与平均意见分的一致性更高，拟合关系也更具线性。

我们还在图 11 中展示了无参考视频质量评价算法在 TV1 上进行 100 次训练测试划分所获得的斯皮尔曼秩相关系数箱线图。与其他算法相比，HDRPatchMAX 的 SRCC 中位数更高，数值离散程度更小。

我们还分别评估了 HDRPatchMAX 中的各个特征集，并在表 V 中报告了 TV1 上的结果。PatchMAX 特征集取得了最高性能，且单独使用时的表现超过了其他无参考视频质量评估算法。我们还对 HDR-PatchMAX 进行了消融研究，逐一移除各个特征集，并评估其余特征集在 TV1 上的表现，结果见表 VI。尽管 PatchMAX 单独使用时表现出色，移除它对算法性能的影响却很小，这表明其他特征集具有稳健性且彼此互补。类似结果

表IV：无参考视频质量评估算法在 LIVE HDRvsSDR 上的 SRCC中位数、LCC 和均方根误差。括号内为标准差。每种电视格式的最佳结果以粗体显示。

数据集	TV1			TV2			TV3		
	斯皮尔曼秩相关系数↑	皮尔逊线性相关系数↑	均方根误差↓	斯皮尔曼秩相关系数↑	皮尔逊线性相关系数↑	均方根误差↓	斯皮尔曼秩相关系数↑	皮尔逊线性相关系数↑	均方根误差↓
RAPIQUE	0.4814(0.1315)	0.5769(0.1099)	17.1953(2.1580)	0.5163(0.1382)	0.5713(0.1332)	17.5951(1.7378)	0.5116(0.1455)	0.5279(0.1402)	16.4634(2.0299)
BRISQUE	0.6923(0.1339)	0.7337(0.1221)	15.6301(2.7714)	0.7355(0.1009)	0.7736(0.0820)	13.2101(2.4794)	0.7285(0.1094)	0.7583(0.0920)	12.9504(2.5819)
TLVQM	0.7340(0.1022)	0.7865(0.0873)	12.9524(2.3336)	0.7422(0.0876)	0.7976(0.0698)	12.4863(2.0532)	0.7734(0.0929)	0.8087(0.0747)	11.7250(1.9711)
ChipQA	0.7396(0.1040)	0.7810(0.0918)	13.0415(2.3269)	0.7450(0.1052)	0.7933(0.0822)	11.9339(1.8891)	0.7943(0.1038)	0.7912(0.0915)	12.2313(2.0476)
HIGRADE	0.7610(0.1196)	0.7928(0.1076)	12.6439(2.7403)	0.7950(0.1189)	0.8246(0.0970)	11.4478(2.3973)	0.7497(0.1074)	0.7679(0.0898)	12.5681(2.2225)
VBLIINDS	0.7895(0.0867)	0.8060(0.0793)	13.4819(2.3147)	0.8005(0.0825)	0.8216(0.0698)	12.0408(2.1253)	0.8052(0.0833)	0.8085(0.0665)	11.2490(2.0545)
VIDEVAL	0.7913(0.0815)	0.8210(0.0722)	11.8524(2.0402)	0.8014(0.0862)	0.8248(0.0702)	11.8614(2.0022)	0.8252(0.0719)	0.8275(0.0648)	11.0899(1.8907)
HDRChipQA	0.7941(0.0954)	0.8129(0.0913)	11.9236(2.4481)	0.8116(0.0774)	0.8209(0.0767)	12.0703(2.2669)	0.8112(0.0899)	0.8135(0.0867)	11.5474(2.3609)
HDRPatchMAX	**0.8586(0.0723)**	**0.8524(0.0654)**	**10.8200(2.0941)**	**0.8436(0.0714)**	**0.8588(0.0636)**	**11.1900(2.0942)**	**0.8494(0.0691)**	**0.8551(0.0627)**	**10.2514(2.0489)**



(a) BRISQUE (b) ChipQA (c) TLVQM (d) VBLIINDS

(e) HIGRADE (f) VIDEVAL (g) HDR ChipQA (h) HDRPatchMAX

图 10：MOS 与无参考视频质量评估预测值的散点图，橙色表示逻辑拟合。

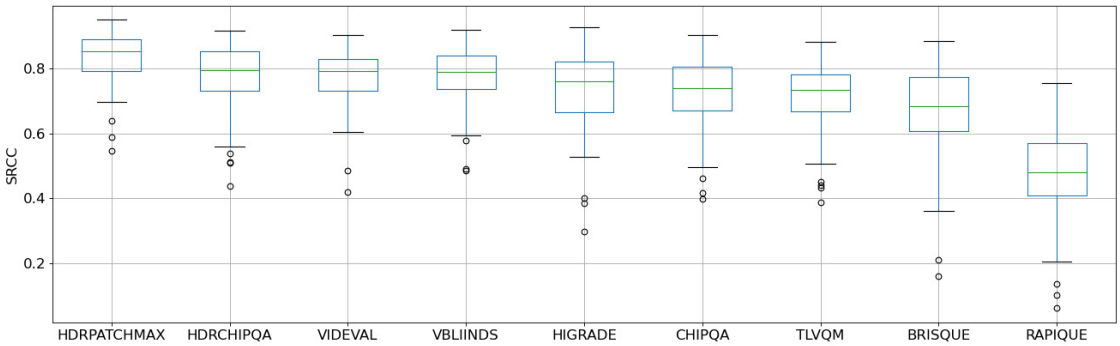


图 11：无参考视频质量评估算法在 100 次训练测试划分中获得的斯皮尔曼秩相关系数箱线图

表V：TV1上单个特征的性能。

特征	斯皮尔曼秩相关系数↑	皮尔逊线性相关系数↑	均方根误差↓
ST梯度芯片	0.5872(0.1380)	0.6264(0.1347)	17.4190(2.5293)
HDRMAX	0.7688(0.0647)	0.7900(0.0682)	13.9499(1.8542)
NIQE	0.7829(0.0896)	0.8057(0.0809)	13.5136(2.3201)
PatchMAX	0.8347(0.0828)	0.8376(0.0794)	12.7068(2.0619)

表六：关于 TV1 的消融研究。

特征集**已移除**	斯皮尔曼秩相关系数↑	皮尔逊线性相关系数↑	均方根误差↓
PatchMAX	0.8553(0.0719)	0.8567(0.0661)	11.9267(2.0255)
ST梯度芯片	0.8524(0.0721)	0.8579(0.0665)	12.0564(1.9838)
NIQE	0.8428(0.0720)	0.8405(0.0715)	12.3666(1.9113)
HDRMAX	0.8334(0.0777)	0.8375(0.0709)	12.5534(1.9670)

在 TV2 和 TV3 上观测到。

在 TV1、TV2 和 TV3 数据上训练的随机森林模型，可以分别视为适用于 OLED 显示屏、QLED 显示屏和 LED 显示屏的独立模型。不过，我们还进行了一项单独的实验：为每台电视分配一个数值（TV1 为 1，TV2 为 2，TV3 为 3），将该数值作为额外特征连接到每种无参考视频质量评价算法上，并使用从所有显示设备获得的分数进行训练；在此过程中，将不同显示设备上显示的视频视为具有不同特征向量的不同刺激。结果如表七所示。HDRPatchMAX 再次取得了最佳表现。

D. 合并数据库上的结果

我们在由 LIVEHDR、LIVE AQHDR 和 LIVEHDRvsSDR 数据库合并而成的数据集上评估了无参考视频质量评估和全参考视频质量评估算法，其中包含 891 个高动态范围视频和 175 个标准动态范围视频。LIVE HDR 和 LIVE AQ HDR 研究使用了 Samsung Q90T （ TV2 ）电视，因此，我们按照第 IV-D 节所述的过程，将 LIVE HDR 和 LIVE AQ HDR 数据库中的分数映射到 LIVE HDRvsSDR 数据库中从 TV2 获得的分数，然后将其用作评估的真实值。全参考指标的结果见表 VIII，无参考指标的结果见表 IX。在合并数据集上，HDRPatchMAX 在无参考视频质量评价模型中表现最佳，而视频多方法评估融合在全参考视频质量评价模型中表现最佳。

表七：使用电视索引作为额外特征时所有显示设备上的结果。

算法	斯皮尔曼秩相关系数↑	皮尔逊线性相关系数↑	均方根误差↓
RAPIQUE	0.5106(0.1306)	0.5789(0.1106)	16.0880(1.9977)
BRISQUE	0.7091(0.1221)	0.7500(0.0926)	13.1656(2.2399)
ChipQA	0.7319(0.1046)	0.7748(0.0884)	12.4513(2.0950)
HIGRADE	0.7673(0.1105)	0.7937(0.0947)	12.3120(2.3021)
TLVQM	0.7491(0.0988)	0.7957(0.0817)	12.0173(2.1908)
HIGRADE	0.7673(0.1105)	0.7937(0.0947)	12.3120(2.3021)
VBLIINDS	0.7850(0.0812)	0.8098(0.0600)	11.4007(1.7672)
HDRChipQA	0.7864(0.0894)	0.8050(0.0878)	11.3890(2.3069)
VIDEVAL	0.7969(0.0812)	0.8243(0.0733)	11.2577(1.9316)
HDRPatchMAX	**0.8367(0.0732)**	**0.8437(0.0605)**	**10.4734(1.9913)**

表 VIII：合并数据库上全参考视频质量评估算法的 SRCC 中位数、线性相关系数和均方根误差。括号内为标准差。最佳结果以粗体显示。

方法	斯皮尔曼秩相关系数↑	皮尔逊线性相关系数↑	均方根误差↓
PSNR	0.3208	0.4777	16.5218
结构相似性 [13]	0.3618	0.3726	17.4523
多尺度结构相似性 [31]	0.3693	0.3794	16.5699
SpEED [15]	0.3235	0.4237	17.0150
STRRED [16]	0.3588	0.3921	17.3007
视频多方法评估融合 [14]	**0.6710**	**0.6930**	**13.5578**
STGREED [17]	0.6182(0.0687)	0.6150(0.0641)	14.8379(1.0856)

表 IX：全参考视频质量评估算法在合并数据库上的 SRCC 中位数、线性相关系数和均方根误差。括号内为标准差。最佳结果以粗体显示。

算法	SRCC↑	皮尔逊线性相关系数↑	均方根误差↓
RAPIQUE	0.5684(0.0828)	0.5440(0.0736)	16.0103(0.9990)
BRISQUE	0.6551(0.0621)	0.6374(0.0560)	14.5969(1.0516)
ChipQA	0.7319(0.1046)	0.7748(0.0884)	12.4513(2.0950)
HIGRADE	0.7673(0.1105)	0.7937(0.0947)	12.3120(2.3021)
TLVQM	0.6997(0.0601)	0.6916(0.0547)	13.7720(1.0247)
HIGRADE	0.6780(0.0736)	0.6670(0.0665)	13.9237(1.2181)
VBLIINDS	0.7501(0.0572)	0.7394(0.0524)	12.8156(1.2017)
VIDEVAL	0.7588(0.0559)	0.7526(0.0529)	12.4759(1.0645)
HDRChipQA	0.7699(0.0537)	0.7702(0.0484)	12.2328(1.1006)
HDRPatchMAX	**0.8064(0.0516)**	**0.8065(0.0445)**	**11.2318(0.9964)**

VI. 结论

我们首次研究了在不同显示设备上比较相同内容、以不同比特率和分辨率编码的高动态范围与标准动态范围视频。我们的研究表明，尽管高动态范围在理论上优于标准动态范围，但其感知质量在实际应用中很大程度上取决于所使用的显示设备。我们还在新数据库上评估了若干无参考视频质量评估和全参考视频质量评估算法，并提出了一种新型无参考视频质量评估算法 HDRPatchMAX；该算法在此数据库上的表现超越了当前最先进水平。我们希望这能推动视频质量评估算法中显示设备建模方面的研究，并促进流媒体传输最佳比特率梯度的研究。

致谢

本研究得到亚马逊公司资助，并得到美国国家科学基金会机器学习基础人工智能研究所（IFML）编号为 2019844 的资助。作者还感谢德克萨斯大学奥斯汀分校的德克萨斯高级计算中心（TACC）提供高性能计算资源，这些资源为本文所报告的研究成果作出了贡献。网址：http://www.tacc.utexas.edu。

参考文献

[1] 国际电信联盟，“BT.709：高清晰度电视标准制作与国际节目交换的参数值”，国际电信联盟，技术报告，2011年。[2] Z. Shang、J. P. Ebenezer、Y. Wu、H. Wei、S. Sethuraman 和 A. C. Bovik，“高运动直播视频的主观与客观质量研究”，*IEEETrans.Image Process.*，第31卷，第1027–1041页，2022年。

[3] D. Y. Lee、S. Paul、C. G. Bampis、H. Ko、J. Kim、S. Y. Jeong、B. Homan 和 A. C. Bovik，“时空子采样视频质量的主观与客观研究”，arXiv预印本 arXiv:2102.00088，2021年。[4] P. C. Madhusudana、X. Yu、N. Birkbeck、Y. Wang、B. Adsumilli 和 A. C. Bovik，“高帧率视频的主观与客观质量评估”，IEEE Access，第9卷，第108 069–108 082页，2021年。[5] R. R. R. Rao、S. G¨oring、W. Robitza、B. Feiten 和 A. Raake，“Avt-vqdb-uhd-1：面向UHD-1的大规模视频质量数据库”，载于2019 IEEE国际多媒体研讨会（ISM）。IEEE，2019年，第17–177页。[6] J. P. Ebenezer、Y. Chen、Y. Wu、H. Wei 和 S. Sethuraman，“低比特率下高运动体育视频的主观与客观质量评估”，载于2022 IEEE国际图像处理会议（ICIP），2022年，第521–525页。[7] V. Hosu、F. Hahn、M. Jenadeleh、H. Lin、H. Men、T. Szir´anyi、S. Li 和 D. Saupe，“康斯坦茨自然视频数据库（Konvid-1k）”，载于国际多媒体体验质量会议，2017年，第1–6页。[8] Y. Wang、S. Inguva 和 B. Adsumilli，“用于视频压缩研究的YouTube UGC数据集”，载于IEEE国际多媒体信号处理研讨会。IEEE，2019年，第1–5页。[9] Z. Ying、M. Mandal、D. Ghadiyaram 和 A. Bovik，“Patch-VQ：‘修补’视频质量问题”，载于IEEE计算机视觉与模式识别会议，2021年，第14 014–14 024页。[10] Z. Shang、J. P. Ebenezer、A. C. Bovik、Y. Wu、H. Wei 和 S. Sethu-raman，“不同环境条件下高动态范围视频的主观评估”，载于IEEE国际图像处理会议，2022年，第786–790页。[11] Z. Shang、J. P. Ebenezer、Y. Wu、H. Wei、S. Sethuraman 和 A. C. Bovik，“高动态范围视频主观与客观质量评估研究”，投稿至 IEEE Trans. Image Process.，2022年。[12] Z. Shang、Y. Chen、Y. Wu、H. Wei 和 S. Sethuraman，“高动态范围体育内容的主观与客观视频质量评估”，载于IEEE/CVF冬季计算机视觉应用会议（WACV）研讨会论文集，2023年1月，第556–564页。[13] Z. Wang、A. Bovik、H. Sheikh 和 E. Simoncelli，“图像质量评估：从误差可见性到结构相似性”，IEEE Trans. Image Process.，第13卷，第4期，第600–612页，2004年。[14] Netflix，《VMAF：旅程仍在继续》，2018年（访问日期：2021年12月28日）。 [在线]。网址：https://netflixtechblog.com/vmaf-the-journey-continues-

44 51 9 12 b ee ed C. G. Bampis、[15] P. Gupta、R. Soundararajan 和 A. C. Bovik，“Speed-qa：用于图像和视频质量的空间高效熵差”，IEEE Signal Process. Letters，第24卷，第9期，第1333–1337页，2017年。[16] R. Soundararajan 和 A. C. Bovik，“通过简化参考时空熵差进行视频质量评估”，IEEE Trans. Circuits Syst. Video Technol.，第23卷，第4期，第684–694页，2012年。[17] P. C. Madhusudana、N. Birkbeck、Y. Wang、B. Adsumilli 和 A. C. Bovik，“St-greed：用于帧率相关视频质量预测的时空广义熵差”，IEEE Trans. Image Process.，第30卷，第7446–7457页，2021年。[18] A. Mittal、A. K. Moorthy 和 A. C. Bovik，“空间域中的无参考图像质量评估”，IEEE Trans. Image Process.，第21卷，第12期，第4695–4708页，2012年。[19] M. A. Saad、A. C. Bovik 和 C. Charrier，“自然视频质量的盲预测”，IEEE Trans. Image Process.，第23卷，第3期，第1352–1365页，2014年。[20] Z. Tu、Y. Wang、N. Birkbeck、B. Adsumilli 和 A. C. Bovik，“UGC-VQA：面向用户生成内容的盲视频质量评估基准”，IEEE Trans. Image Process.，第30卷，第4449–4464页，2021年。[21] Z. Tu、X. Yu、Y. Wang、N. Birkbeck、B. Adsumilli 和 A. C. Bovik，“RAPIQUE：用户生成内容的快速准确视频质量预测”，arXiv预印本 arXiv:2101.10955，2021年。[22] J. P. Ebenezer、Z. Shang、Y. Wu、H. Wei、S. Sethuraman 和 A. C. Bovik，“ChipQA：通过时空芯片进行无参考视频质量预测”，IEEE Trans. Image Process.，第30卷，第8059–8074页，2021年。[23] ——，“HDR-ChipQA：高动态范围视频的无参考质量评估”，投稿至 IEEE Trans. Image Process.，2023年。[24] A. Mittal、R. Soundararajan 和 A. C. Bovik，“构建一个‘完全盲’图像质量分析器”，IEEE Signal Process. Lett.，第20卷，第3期，第209–212页，2012年。[25] J. Korhonen，“无参考消费级视频质量评估的两级方法”，IEEE Trans. Image Process.，第28卷，第12期，第5923–5938页，2019年。[26] 国际电信联盟，“BT.2020：超高清晰度电视系统制作与国际节目交换的参数值”，国际电信联盟，技术报告，2015年。[27] S. Miller、M. Nezamabadi 和 S. Daly，“用于更高效使用比特编码的感知信号编码”，*SMPTEMotion Imaging J.*，第122卷，第4期，第52–59页，2013年。[28] Rtings，《峰值亮度测量》，2023年（访问日期：2023年2月28日）。[在线]。可访问：https://www.rtings.com/tv/tests/picture-quality/sdr-peak-brightness [29] 国际电信联盟，“BT.500：电视图像质量主观评价方法”，国际电信联盟，技术报告，2019年。[30] ——，“ITU. 910：多媒体应用的主观视频质量评价方法”，国际电信联盟，技术报告，2008年。[31] Z. Wang、E. Simoncelli 和 A. Bovik，“用于图像质量评估的多尺度结构相似性”，载于*Asilomar Conf. Signals, Syst., Comput.*，第2卷，2003年，第1398–1402页，第2卷。[32] H. R. Sheikh、M. F. Sabir 和 A. C. Bovik，“近期全参考图像质量评估算法的统计评估”，*IEEE Trans. Image Process.*，第15卷，第11期，第3440–3451页，2006年。[33] J. P. Ebenezer、Z. Shang、Y. Wu、H. Wei、S. Sethuraman 和 A. C. Bovik，“使视频质量评估模型对位深度具有鲁棒性”，已投稿至 *IEEE Signal Process. Letters*，2023年。[34] J. P. Ebenezer、Z. Shang、Y. Wu、H. Wei 和 A. C. Bovik，“使用时空芯片进行无参考视频质量评估”，载于*IEEE Intl. Workshop Multimedia Signal Process.*，2020年，第1–6页。[35] D. W. Dong 和 J. J. Atick，“时间去相关：外侧膝状体中滞后与非滞后响应的理论”，*Netw.: Comput. Neural Syst.*，第6卷，第2期，第159–178页，1995年。