arXiv:2301.01069v1 [eess.IV] 3 Jan 2023

面向压缩视频质量评估的显著性感知时空伪影检测

Liqun Lin、Yang Zheng、Weiling Chen、Chengdong Lan、Tiesong Zhao

摘要——**压缩视频常会出现令人不适的视觉伪影，即可感知编码伪影（PEA），它们会严重降低视频视觉质量。能够识别和量化各种 PEA 的主观与客观测量方法，对于提升视觉质量至关重要。本文研究了四种空间 PEA（即模糊、块效应、色彩渗漏和振铃）以及两种时间 PEA（即 闪烁和漂浮）对视频质量的影响。针对空间伪影，我们提出了一种计算成本低、且与人类视觉感知一致性更高的视觉显著性模型。针对时间伪影，我们改进了基于自注意力的 TimeSFormer，以检测时间伪影。基于这六种 PEA，我们提出了一种名为显著性感知时空伪影测量（SSTAM）的质量指标。实验结果表明，所提方法优于当前最先进的指标。我们认为，SSTAM 将有助于优化视频编码技术。**

索引词—**视频质量评估、显著性检测、可感知编码伪影（PEA）、压缩伪影。**

I. 引言

WITH 随着用户需求不断增长，高清（HD）/超高清（UHD）视频得到了广泛应用。由于数据量庞大，HD/UHD 视频需要经过编码，以满足有限的传输带宽和存储空间要求。编码后的视频会出现令人不适的可感知编码伪影（PEA），从而导致视觉质量下降 [1]。因此，有必要基于 PEA 开展视频质量评估（VQA），以指导视频编码优化、减少失真并提升用户的视觉体验。

视频质量评估（VQA）可分为主观方法和客观方法。对于主观视频质量评估，国际电信联盟（ITU） [2]提出的平均意见分（MOS）和差异平均意见分（DMOS）用于反映视频的主观质量。主观质量评估准确可靠，但耗时且劳动强度大。

客观视频质量评估方法有三种：全参考（FR）、简化参考（RR）和无参考

本研究主要得到国家自然科学基金资助，项目编号为 62171134 和 61901119。此外，本研究还得到福建省自然科学基金资助，项目编号为 2022J02015 和 2022J05117。（通讯作者：C. Lan。）

L. Zhao 任职于福州大学物理与信息工程学院、福建省媒体信息智能处理与无线传输重点实验室（中国福建省福州市，邮编 350116），同时也就职于中国福建光电信息科学与技术创新实验室（电子邮箱： {lin liqun, t.zhao}@fzu.edu.cn）。

Y. Zheng、W. Chen、C. Lan 任职于福州大学物理与信息工程学院福建省媒体信息智能处理与无线传输重点实验室，地址为中国福建省福州市 350116（电子邮箱： {211127117, weiling.chen、lancd}@fzu.edu.cn）。

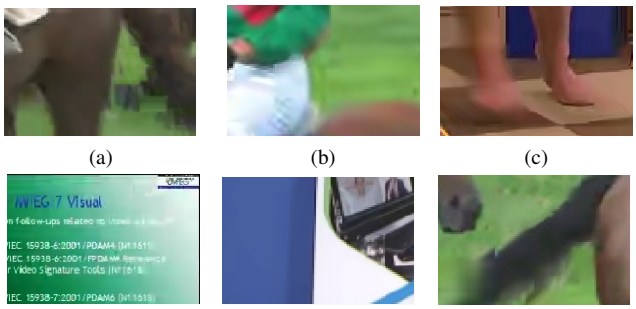


图 1. PEA 示例。(a) 块效应：块效应：块效应：常用编码器采用基于块的处理方式，所有压缩过程都在互不重叠的块内进行。这通常会导致块边界处出现伪不连续现象。块效应的视觉表现可能因视觉不连续区域而异[3]。(b) 模糊：模糊：模糊：视频信号重建可能会造成高频能量显著损失，从而导致视觉模糊。从感知角度来看，模糊通常表现为图像空间细节丢失，或边缘、纹理区域的清晰度下降 [4]。(c) 颜色渗漏：颜色渗漏：颜色渗漏：当视频中一种颜色的边缘无意间溢出或覆盖到另一种颜色上时，就会出现这种现象[5]。(d) 振铃：振铃：振铃：它会在锐利边缘附近呈现“光晕”“环纹”或“阴影”的形式 [6]。(e) 闪烁：闪烁：闪烁：指亮度或色度沿时间维度频繁变化。这种现象十分醒目，也令观看者感到厌烦。(f) 漂浮：漂浮：漂浮：指某些区域产生错觉般的移动，而周围区域保持静止。从视觉上看，这些区域似乎漂浮在周围背景之上。

（NR） [7]。全参考视频质量评估方法通过计算原始视频与压缩视频之间的差异来衡量视频质量。一些全参考方法，如峰值信噪比（PSNR）和结构相似性（SSIM）[8]，得到了广泛应用。部分参考方法利用原始视频的少量特征来评估压缩视频的质量。时空部分参考熵差异（STRRED） [9]和空间有效熵差异质量评估（SpEED-QA） [10] 等部分参考视频质量评估指标也表现良好。然而，这些方法需要原始视频数据，而这类数据通常无法在视频质量评估中获取，从而降低了模型的泛化能力。

最受欢迎的质量评估方案是无参考视频质量评估（NR-VQA），因为终端用户无法获取原始视频。一些无参考视频质量评估方法，例如视频内在完整性与失真评估预言机（VIIDEO） [11]和双层视频质量模型（TLVQM） [12]，从不同角度提升了视频质量评估的性能。最近，人们更加关注压缩伪影对压缩视频质量的影响。例如，Bovik[13] 考虑了视频编码伪影，但未进行具体分类。在 [6], 中，研究人员通过利用振铃和块效应伪影开发了一种无参考视频质量评估方法。TLVQM [12]和感知型伪影测量（SAAM）中考虑了两种以上的空间PEA，而显著性-

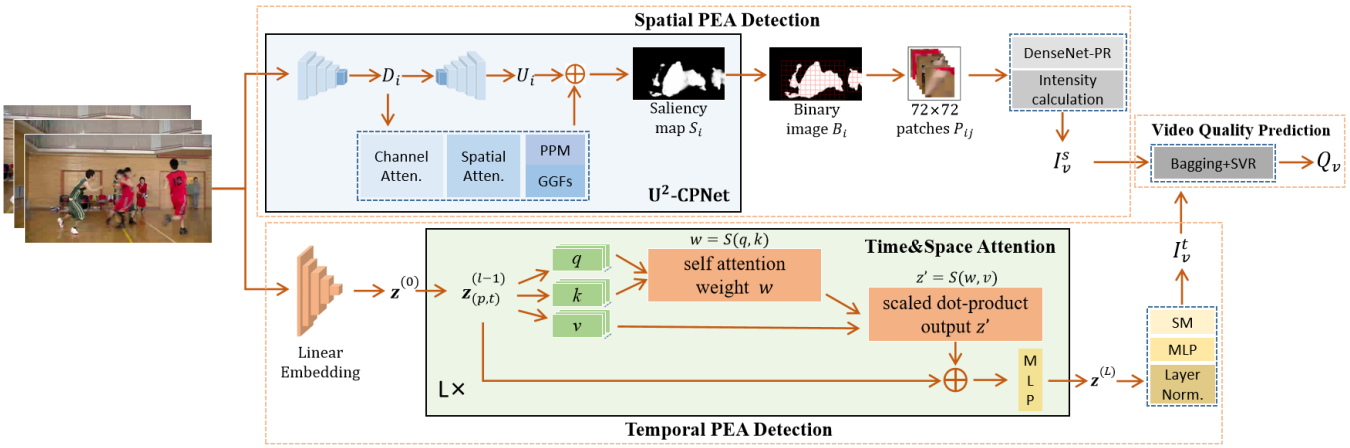


图 2. SSTAM 的整体框架。在空间 PEA 检测中，每个视频帧依次经过空间 PEA 模块，以检测空间 PEA 的强度 I_v^s。在检测时间 PEA 时， $z^{(l-1)}$ 依次经过 Transformer 编码器块 $l \in [1, L]$ 中的时间注意力模块、空间注意力模块和多层感知机。时间 PEA 模块的输出为 I_v^t。

感知型伪影测量（SAAM） [14]。文献 [15], 提到了模糊、块效应和闪烁伪影对视频质量的影响，但并未进行具体检测。

如上所述，已有若干视频质量评估方法考虑了压缩伪影，但其中大多数仅涉及少数几种压缩伪影。此外，这些方法的计算开销非常大 [14]。为解决这些问题，我们提出了一种名为显著性感知时空伪影测量（SSTAM）的方法。本文考虑了六种对压缩视频质量有重大影响的典型 PEA, [16], 如图 1 所示。主要贡献总结如下。

(1)提出了一种基于 PEA 检测的无参考视频质量评估方法。PEA 检测模块能够准确识别六种典型的 PEA。基于该模块，可获取 PEA 强度，用于预测视频质量。

(2)提出了一种利用视觉显著性、使结果与人类视觉感知高度一致的显著性模型。该模型使我们的方法更适用于压缩视频的质量评估，同时排除对视频视觉质量影响甚微的区域，从而降低计算开销。

(3)在四个公开数据库上进行了全面验证，以检验所提出的 SSTAM。结果表明，SSTAM 在视频质量评估方面表现出良好的性能。

II. 基于 PEA 的视频质量指标

本节系统地介绍 SSTAM 的框架，其整体结构如图 2 所示。SSTAM 由空间 PEA 检测模型、时间 PEA 检测模型和视频质量预测模型组成。

A. 空间 *PEA* 检测模型

空间 PEA 检测模型可分为视觉显著性模型和空间 PEA 检测两部分。首先，视觉显著性模型提取显著性区域。该模型可以排除对视觉体验影响甚微的区域，使其更符合人类视觉感知，同时降低计算复杂度。随后，空间 PEA 检测部分会获取显著性区域内的空间 PEA。

视觉显著性模型：视觉显著性模型：视觉显著性模型：视觉显著性是人类视觉系统（HVS）的重要组成部分，也是决定视频感知质量的因素之一。显著性目标检测用于识别图像或视频中视觉上最突出的目标。聚焦于显著性区域还能显著降低网络的计算负担。现有显著性模型主要专注于提取特定类别的目标 [17]–[20]。对于视频，目标运动和外观对比度的变化更能吸引注意力，因此提取的显著性目标应更多地聚焦于此类区域 [17]。鉴于压缩伪影的特点，我们更关注通道注意力，也称为特征注意力。通道关联用于体现各项特征的重要性。此外，我们还采用空间注意力来补充通道注意力。本研究使用卷积块注意力模块（CBAM）提取通道和空间特征 [21]。

然而，要使网络更好地识别显著性区域，将提取的通道特征和空间特征与语义信息相结合至关重要。尽管高层语义特征有助于定位显著目标，较低层级的语义特征也能提供必要的局部细节。在自顶向下过程中，低层语义信息往往会被稀释。局部信息的缺失常常会干扰显著目标的定位。受 U-Net [22] 和 U²-Net [23], 的启发，我们提取多层级特征并将其融合，以实现更好的检测。为妥善利用各层级的信息，我们采用全局引导流（GGFs） [24]来确定自顶向下路径各部分中全局引导信息的权重。它能够将金字塔池化模块（PPM） [24] 生成的信息与自顶向下通路输出的特征图相结合。受此启发，我们提出了一种改进的网络——U²-卷积金字塔网络（U²-CPNet），将注意力机制与全局引导相结合。

此外，还改进了损失函数，使

模型更有针对性地识别显著性区域。在训练过程中，二元交叉熵（BCE）损失是像素级的，它不考虑相邻区域的标签，并对显著性区域和背景中的像素赋予相同的权重。该损失函数旨在使所有像素都收敛，因此不太适合提取显著性区域的目标。训练时，损失函数应更多关注相邻像素之间的相关性以及全局层面。本文使用 BCE 函数，使网络能够充分获取像素级数据。分别使用交并比（IoU）和结构相似性（SSIM）处理全局背景以及相邻像素之间的局部信息。IoU 使网络能够更多地关注显著性图的全局上下文，并聚焦于前景目标[25]。BCE 和 IoU 损失函数的表达式见公式 1 和公式 2。

$$L_{\text{bce}} = -(1-G_{(a,b)})\log(1-S_{(a,b)}) - G_{(a,b)}\log S_{(a,b)}, \quad (1)$$

$$L_{\text{iou}} = \frac{\sum_{a=1}^{H}\sum_{b=1}^{W} S_{(a,b)}G_{(a,b)}}{\sum_{a=1}^{H}\sum_{b=1}^{W}[S_{(a,b)} + G_{(a,b)} - S_{(a,b)}G_{(a,b)}]}, \quad (2)$$

其中， a 和 b 表示每个视频帧中像素的垂直和水平坐标。 H 和 W 表示视频帧的高度和宽度。 $G_{(a,b)} \in \{0,1\}$ 表示像素（a， b）的真实标签， $S_{(a,b)} \in \{0,1\}$ 是显著目标的预测概率。

结构相似性用于处理图像内容中相邻像素之间的结构关系。我们将其融入训练损失，以突出显著目标的结构信息。这确保了有足够的空间进行特征利用。当显著性检测区域中的概率降为零时，背景在视觉上会显得干净得多。结构相似性损失函数的表达式如式 3 所示。

$$L_{\text{ssim}} = 1 - \frac{(2\mu_G\mu_P + C_1^2)(2\sigma_{GP} + C_2^2)}{(\mu_G^2 + \mu_P^2 + C_1^2)(\sigma_G^2 + \sigma_P^2 + C_2^2)}, \quad (3)$$

其中，下标 G 表示二值真实掩码，下标 P 表示预测概率图掩码。 μ_G 和 μ_P 指均值， σ_G 和 σ_P 表示二者的标准差， σ_{GP} 表示二者的协方差。 C_1 和 C_2 分别为 0.01 和 0.03 [26], 。它们用于避免除零错误。

模型训练过程中混合损失的策略如公式 4 所示，其中 L 是训练过程的损失，且每个组成部分的权重均为 1。

$$L = L_{\text{bce}} + L_{\text{iou}} + L_{\text{ssim}}. \quad (4)$$

空间 PEA 检测：空间 PEA 检测：空间 PEA 检测：对于空间 PEA 识别，用于 PEA 识别的 DenseNet（ DenseNet-PR） [16] 可以缓解梯度消失问题并改善特征传递，同时大幅减少参数数量。对视频帧进行预处理，使其符合 DenseNet-PR 的输入要求：获得由 U²-CPNet 提取的显著性图 S_i 后，将其二值化为 B_i，然后将 B_i 映射回视频帧，并将视频帧中的显著性区域划分为 72×72 图像块。训练四种空间 PEA 识别模型，以检测图像块中的 PEA。由此获得每个 72×72 图像块的概率列表，即 I_{ij}。最后，通过以下公式计算整个视频序列中每种空间 PEA 的强度。

$$I_v^s = \frac{1}{N_f N_p}\sum_{i=1}^{N_f}\sum_{j=1}^{N_p} I_{\text{ij}}, \quad (5)$$

其中， N_p 表示每帧显著性区域中的图像块总数。 N_f 表示视频帧总数。 I_v^s 表示每个视频中空间 PEA 的强度值。 I_{ij}如上所述，表示每个 72×72 图像块中某种 PEA 的概率。 i 是视频的第 i帧， j 是视频帧中的第 j个图像块。

B. 时间*PEA*检测模型

特定于视频的任务同时考虑空间和时间信息。与其他视频理解任务相比，时间 PEA 检测更加依赖时空特征，而网络捕捉这些特征的难度更大。同时，目前关于时间 PEA 分析与检测的研究非常匮乏。因此，开展时间 PEA 检测至关重要。自注意力机制能够通过直接比较所有时空特征，捕捉长距离范围内的局部和全局依赖关系。因此，我们改进了基于自注意力的 TimeSFormer [27]，并将其应用于时间 PEA 检测。由于时间 PEA 的空间特征相对不显著，我们在训练网络时更加关注时间特征。

闪烁和漂浮的训练数据源自数据库 PEA265 [16]。时间 PEA 检测模型如图 2 所示。每个视频的时间 PEA 强度最终计算如下。

$$I_v^t = \text{SoftMax}(\text{MLP}(\text{LN}(\sum_{j=1}^{N_c}\sum_{i=1}^{N_t}\text{TSF}_j(F_i)))), \quad (6)$$

其中，我们将连续的 N_t 帧划分为一个片段，N_c 是每个视频的片段总数。TSF() 表示 TimeSFormer 的处理过程，LN() 表示层归一化。I_v^t 表示每个视频中的时间 PEA 强度。

C. 视频质量预测

预测视频质量需要有效地映射前文检测到的空间和时间伪影与主观质量分数之间的相关性。我们在集成学习中采用自助聚合（装袋法） [28] 方法。使用训练集不同部分的多个基学习器相结合，能够比单个学习器取得更好的泛化性能。假设有一个完整的数据集 D，我们将 MOS|DMOS 值与 D 中每个视频的六个 PEA 的强度相匹配。匹配过程如下式所示。

$$D = \{(I_{V_1}, \text{MOS}_1|\text{DMOS}_1), \ldots, (I_{V_k}, \text{MOS}_m|\text{DMOS}_m)\}, \quad (7)$$

其中，MOS_m|$DMOS_m$ 表示第 m 个压缩视频的主观质量分数，I_{V_k} 表示所有空间和时间感知显著性区域的强度汇总值。

训练时，我们按 8:2 的比例将数据集划分为训练集和测试集，然后将训练集均分为十个子训练集。使用这十个子训练集训练支持向量回归（SVR）模型。通过预测质量分数与真实质量分数之间的皮尔逊线性相关系数（PLCC）评估这些基学习器的性能。从十个基学习器中选出 PLCC 分数最高的三个学习器来预测视频质量 Q_v。Q_v 的表达式如下所示。

$$Q_v = \sum_{i=1}^{n} \omega_i y_i(x), \tag{8}$$

$$\sum_{i=1}^{n} \omega_i = 1 \quad \omega_i \in [0,1), \tag{9}$$

其中，x 是测试集。$y_i(x)$表示第 i个基学习器的预测输出。ω_i 是第 i个基学习器的权重。n 表示基学习器的数量，其值为 10。

III. 实验与讨论

我们在四个常用且可公开获取的视频质量数据库（VQDs）上开展实验与评估，包括 LIVE [29], CSIQ [30], IVP [31],和 FERIT-RTRK [32], ，以验证 SSTAM 的泛化能力。LIVE 包含 40 个分辨率为768×432的压缩视频。CSIQ 包含 36 个分辨率为 832×480的压缩视频。IVP 和 FERIT-RTRK 分别包含 40 个和 30 个分辨率为 1920×1080 的压缩视频。在本研究中，我们将这四个数据库中压缩视频的六种压缩伪影强度与其 MOS|DMOS 相匹配，构建了四个完整的数据集用于评估。我们将 SSTAM 与若干经典且广受欢迎的质量评估方法进行比较，以验证其性能。这些方法包括 PSNR、SSIM [8], 多尺度结构相似性[33], 时空偶然失真的降低 [9], SpEED-QA [10], VIIDEO [11],TLVQM[12], SAAM [14], BRISQUE [34]和 NIQE [35]。采用 PLCC 和斯皮尔曼等级相关系数（SRCC）值来衡量这些方法的性能。

表 I 以皮尔逊线性相关系数（PLCC）为指标的性能比较。

方法	LIVE	CSIQ	IVP	FERIT-RTRK	总体
PSNR	0.5735	0.8220	0.7998	0.7756	0.7383
SSIM	0.6072	0.8454	0.8197	0.6870	0.7406
多尺度结构相似性	0.6855	0.8782	0.8282	0.8724	0.8105
时空偶然失真的降低	0.8392	0.8772	0.5947	0.8425	0.7823
SpEED-QA	0.7933	0.8554	0.6822	0.6978	0.7586
BRISQUE	0.2154	0.5526	0.2956	0.7653	0.4335
NIQE	0.3311	0.5350	0.3955	0.5817	0.4505
VIIDEO	0.6829	0.7211	0.4358	0.3933	0.5651
TLVQM	0.7511	0.7740	-	-	0.7619
SAAM	0.9023	0.9244	0.8717	0.9499	0.9091
SSTAM	0 94900 94900...9490 0.95230 95230..9523 0.94780 94780 9478..				

表 II SRCC 方面的性能比较。

方法	LIVE	CSIQ	IVP	FERIT-RTRK	总体
PSNR	0.4146	0.8028	0.8154	0.7685	0.6928
SSIM	0.5677	0.8440	0.8049	0.7236	0.7328
多尺度结构相似性	0.6773	0.9465	0.7917	0.8508	0.8107
时空偶然失真的降低	0.8358	0.9770097700..9770	0.8595	0.8310	0.8761
SpEED-QA	0.7895	0.9639	0.8812	0.7945	0.8587
BRISQUE	0.2638	0.5655	0.1051	0.7574	0.3961
NIQE	0.1769	0.5012	0.2351	0.4855	0.3362
VIIDEO	0.6593	0.7153	0.1621	0.3177	0.4667
TLVQM	0.7338	0.7956	-	-	0.7631
SAAM	0.8691	0.8810	0.8413	0.9429	0.8796
SSTAM	09201092010...9201 0.9177 0.90480904809048..				

表 III 消融实验。

指标	SAAM	SAAM2.0	SSTAM
PLCC	0.9091	0.9475	0.9552
SRCC	0.8796	0.9109	0.9281

性能比较：性能比较：性能比较：结果见表 I 和表 II。加粗的结果代表各数据库中表现最佳的方法。显然，我们的模型在这四个数据库上都极具竞争力。根据各数据库中压缩视频的数量加权，得出总体性能。如最后一列所示，SSTAM 的整体性能优于其他方法。与其他方法相比，SAAM 和 SSTAM 的显著提升，是因为它们充分考虑了 PEA 对压缩视频质量的影响。这项实验也充分证明，PEA 与压缩视频的主观质量分数之间存在很强的相关性。我们发现，在 FERIT-RTRK 中，包含显著性模型的方法对视频质量的预测更为有效。由于 FERIT-RTRK 中的显著性目标相对较少且尺寸较大，网络能够更准确地提取视频中的显著性区域

。因此，更好地提取显著性区域能够显著提升质量评估的性能。

消融实验：消融实验：消融实验：为进一步验证所提出策略对视频质量评估的提升作用，我们分别比较了显著性模型和时间 PEA 的影响。SAAM 中采用了名为 AttentiveCNN-LSTM 网络（ACLNet）的视觉显著性模型 [36] 。为评估性能，我们用 U^2- CPNet 替换 SAAM 中的 ACLNet。更换显著性网络后，我们将该方法称为 SAAM2.0。如表 III 所示，我们的显著性模型在识别显著性区域方面表现更好，并提升了网络的性能。SAAM2.0 与 SSTAM 的对比表明，时间 PEA 确实会产生影响。

IV. 结论

本文提出了一种基于显著性感知时空伪影检测的压缩视频质量指数。实验结果表明，所提出的 SSTAM 在性能和泛化能力方面均有所提升。我们相信，未来它将在压缩视频质量评估和视频编码优化中发挥重要作用。

参考文献

[1] J. Xia、Y. Shi、K. Teunissen 和 I. Heynderickx，“压缩视频中可感知的伪影及其与视频质量的关系”，载于《信号处理：图像通信》，第24卷，第7期，第548–556页，2009年。[2] 国际电信联盟。《电视图像主观质量评估方法》；ITU-R BT.500-13号建议书；国际电信联盟：日内瓦，瑞士，2012年。[3] B. Xie、H. Zhang 和 C. Jung，“WCDGAN：用于去除高度压缩图像伪影的弱连接密集生成对抗网络”，载于《IEEE Access》，第10卷，第1637-1649页，2022年。[4] X. Sheng、L. Li、D. Liu 和 Z. Xiong，“基于几何的点云压缩中的属性伪影去除”，载于《IEEE图像处理汇刊》，第31卷，第3399-3413页，2022年。[5] Q. Zhang、Y. Nie、L. Zhu、C. Xiao 和 W. -S. Zheng，“用于忠实重着色的盲颜色分离模型”，载于《IEEE多媒体汇刊》，第24卷，第1545–1557页，2022年。[6] M. K. Rohil、N. Gupta 和 P. Yadav，“一种改进的无参考图像质量评估模型及一种基于帧分析的无参考视频质量评估模型”，载于《信号、图像与视频处理》，第14卷，第1期，第205–213页，2020年。[7] J. Vlaovi´c、M. Vranje˘s、D. Grabi´c 和 D. Samard˘zija，“不同空间分辨率视频的客观视频质量评估方法比较”，载于国际系统、信号与图像处理会议（IWSSIP），第287–292页，2019年。[8] Z. Wang、A. C. Bovik、H. R. Sheikh 和 E. P. Simoncelli，“图像质量评估：从误差可见性到结构相似性”，载于《IEEE图像处理汇刊》，第13卷，第4期，第600–612页，2004年。[9] R. Soundararajan 和 A. C. Bovik，“基于部分参考时空熵差分的视频质量评估”，载于《IEEE视频技术电路与系统汇刊》，第23卷，第4期，第684–694页，2013年。[10] C. G. Bampis、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分”，载于《IEEE信号处理快报》，第24卷，第9期，第1333–1337页，2017年。[11] A. Mittal、M. A. Saad 和 A. C. Bovik，“完全盲的视频完整性判定器”，载于《IEEE图像处理汇刊》，第25卷，第1期，第289–300页，2016年。[12] J. Korhonen，“无参考消费级视频质量评估的两级方法”，载于《IEEE图像处理汇刊》，第28卷，第12期，第5923–5938页，2019年。[13] T. R. Goodall 和 A. C. Bovik，“视频失真的检测与映射”，载于《IEEE图像处理汇刊》，第28卷，第6期，第2680–2691页，2018年。[14] L. Lin、Y. Jing、W. Zheng 等，“基于显著性感知伪影检测的压缩视频质量指数”，载于《传感器》，第21卷，第19期，2021年，文章编号6429。[15] C. Jin、Z. Peng、F. Chen 和 G. Jiang，“窗口化6DoF合成视频的主客观视频质量评估”，载于《IEEE广播汇刊》，第68卷，第3期，第594–608页，2022年。[16] L. Lin、S. Yu、L. Zhou、W. Chen、T. Zhao 和 Z. Wang，“PEA265：视频压缩伪影的感知评估”，载于《IEEE视频技术电路与系统汇刊》，第30卷，第11期，第3898–3910页，2020年。[17] Y. Peng、L. Guan、X. Yuan 等，“使用伪标签的半监督视频显著目标检测”，载于IEEE/CVF国际计算机视觉会议（ICCV）论文集，第7284–7293页，2019年。[18] J. Guo、Y. Luo，“基于生成对抗网络的无参考全景视频质量评估”，载于《多媒体工具与应用》，第80卷，第18期，第27531–27552页，2021年。[19] Y. Zhao、J. Zhao、J. Li 和 X. Chen，“具有全方位目标感知能力的RGB-D显著目标检测”，载于《IEEE图像处理汇刊》，第30卷，第7717–7731页，2021年。[20] N. Liu、N. Zhang、K. Wan、L. Shao 和 J. Han，“视觉显著性变换器”，载于IEEE/CVF国际计算机视觉会议（ICCV）论文集，第4702–4712页，2021年。[21] W. Sanghyun、P. Jong、L. Joon 等，“CBAM：卷积块注意力模块”，载于欧洲计算机视觉会议（ECCV）论文集，第3–19页，2018年。[22] O. Ronneberger、P. Fischer 和 T. Brox，“U-Net：用于生物医学图像分割的卷积网络”，载于医学图像计算与计算机辅助干预国际会议（MICCAI），第234–241页

[23] Q. Xue、Z. Zi、H. Chen 等，“U2-Net：采用嵌套 U 结构深入探索显著性目标检测”，载于模式识别，第 106 卷，第 107–404 页，2020年。[24] L. Jiang、H. Qi、C. Ming 等，“一种用于实时显著性目标检测的简易基于池化的设计”，载于*IEEE/CVF*计算机视觉与模式识别会议 *(CVPR)*，第 3917–3926 页，2019年。[25] M. Rahman 和 Y. Wang，“在深度神经网络中优化图像分割的交并比”，载于国际视觉计算研讨会 *(ISVC)*，第 234–244 页，2016年。[26] Q. Xue、Z. Zi、H. Chen 等，“Basnet：边界感知显著性目标检测”，载于*IEEE/CVF*计算机视觉与模式识别会议 *(CVPR)*，第 7479-7489 页，2019年。[27] B. Gedas、H. Wang 和 L. Torresani，“时空注意力是否就是视频理解所需的一切”，载于国际机器学习会议 *(ICML)*，第 2 卷，第 3 期，第 4–6 页，2021年。[28] L. Breiman，“装袋预测器”，载于机器学习，第 24 卷，第 123–140 页，1996年。[29] S. Hu、L. Jin、H. Wang、Y. Zhang、S. Kwong 和 C.-C. J. Kuo，“基于感知加权均方误差的客观视频质量评估”，载于*IEEE*视频技术电路与系统汇刊，第 27 卷，第 1844–1855 页，2017年。[30] P. V. Vu、D. M. Chandler，“ViS3：一种通过分析空间和时空切片进行视频质量评估的算法”，载于电子成像期刊，第 23 卷，第 13–16 页，2014年。[31] F. Zhang、S. Li、L. Ma、Y. Wong 和 K. Ngan，“IVP主观视频质量数据库”，在线资源：http://ivp.ee.cuhk.edu.hk/research/database/subjective，2021年。[32] V. Baj˘ cinovci、M. Vranje˘s、D.Babi´ c 和 B.Kova˘ cevi´ c，“MPEG-2、H.264 和 H.265 视频的主观与客观质量评估”，载于*2017*年*ELMAR*国际研讨会，第 73–77 页，2017年。[33] W. Sun、Q. Liao、J. Xue 和 F. Zhou，“SPSIM：一种用于全参考图像质量评估的基于超像素的相似性指标”，载于*IEEE*图像处理汇刊，第 27 卷，第 9 期，第 4232–4244 页，2018年。[34] A. Mittal、A. K. Moorthy 和 A. C. Bovik，“空间域中的无参考图像质量评估”，载于*IEEE*图像处理汇刊，第 21 卷，第 12 期，第 4695–4708 页，2012年。[35] A. Mittal、R. Soundararajan 和 A. C. Bovik，“制作一个‘完全盲’的图像质量分析器”，载于*IEEE*信号处理快报，第 20 卷，第 3 期，第 209–212 页，2013年。[36] W. Wang、J. Shen、J. Xie 等，“深度学习时代重新审视视频显著性预测”，载于*IEEE*模式分析与机器智能汇刊，第 43 卷，第 1 期，第 220–237 页，2019年。

，2015年。